Ph.D. DISSERTATION

Hierarchical Trajectory Matching for Wide-Area Multi-Pedestrian Tracking

광역 다중 보행자 추적을 위한 계층적 궤적매칭 기법

BY

Kikyung Kim

AUGUST 2020

DEPARTMENT OF ELECTRICAL ENGINEERING AND
COMPUTER SCIENCE
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

Hierarchical Trajectory Matching for Wide-Area Multi-Pedestrian Tracking

광역 다중 보행자 추적을 위한 계층적 궤적매칭 기법

지도교수 최 진 영

이 논문을 공학박사 학위논문으로 제출함

2020 년 8 월

서울대학교 대학원

전기 컴퓨터 공학부

김 기 경

김 기 경의 공학박사 학위논문을 인준함

2020 년 8 월

위 원 장	고 형 석	
부위원장	최 진 영	
위 원	조 남 익	
위 원	정 교 민	
위 원	강 훈	

Abstract

The purpose of wide-area tracking problem is to track pedestrians that appear on cameras that overlap or do not overlap, regardless of the time interval or person density. In a single camera tracking, data association using overlapping of the detection boxes is used to solve the tracking problem, but still has appearance ambiguity issues. However, wide-area tracking requires a tracking scheme that focuses on the appearance similarity of humans, without the use of overlapping of detection boxes. In this dissertation, we propose the tracking scheme for the Wide-area Multi-Pedestrian Tracking (WaMuPeT). To achieve the WaMuPeT, we propose the trajectory matching in overlapping camera settings (Ch. 3), non-overlapping camera settings (Ch. 4) and robust trajectory matching in dense scene settings (Ch. 5).

In trajectory matching in overlapping camera settings (Ch. 3), we propose a novel deep-learning architecture for accurate 3-D localization and tracking of a pedestrian using multiple cameras. The deep-learning network is composed of two networks: detection network and localization network. The detection network yields the pedestrian detections and the localization network estimates the ground position of a pedestrian within its detection box. In addition, an attentional pass filter is introduced to effectively connect the two networks. Using the detection proposals and their 2-D grounding positions obtained from the two networks, multi-camera multi-target 3-D localization and tracking algorithm is developed through min-cost network flow approach. In the experiments, it is shown that the proposed method improves the performance of 3-D localization and tracking.

In trajectory matching in non-overlapping camera settings (Ch. 4), we propose a novel re-ranking method using a ranking-reflected metric to measure the similarity

between two ordered sets of K-nearest neighbors (OKNN). The proposed metric for ranking-reflected similarity (RSS) reflects the ranking of the shared elements between the two OKNNs. Using RSS, a re-ranking procedure is proposed that prioritizes galleries having neighbors similar to a probe's neighbor in the perspective of ranking order. In the experiment, we show that the proposed method improves the Re-ID accuracy by add-on to the state-of-the-art methods.

In robust trajectory matching in dense scene settings (Ch. 5), we propose a novel framework for multi-pedestrian tracking to generate robust trajectories in dense scene. In the proposed tracking method, we propose the tracking method based on the trajectory matching by the strategy of divide and conquer method. In this strategy, short-term, mid-term and long-term trajectories are generated by each trajectory merging stages, respectively. Also we propose a novel deep-feature matching method called stable boundary selection (SBS). In SBS matching, the detections are clustered by the group similarity of deep features, so that robust trajectories can be generated. With the smoothing algorithms and the detection restoration algorithm, the proposed tracking method shows the state-of-the-art tracking accuracy in three public tracking dataset.

Keywords: wide-area tracking, multi-pedestrian tracking, pedestrian detection, pedestrian localization, person re-identification, re-ranking

Student Number: 2013-20748

Contents

List of Figures

List of Tables

Chapter 1

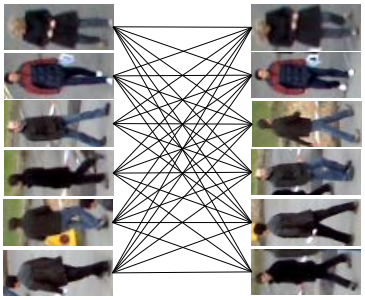
Introduction

1.1 Background

With advances in computer learning technology, computer scientists are attempting to apply computer vision technology to many aspects of the industry. Among them, computer vision applications targeting pedestrians (we call this the pedestrian visions in this dissertation) have attracted the attention of many researchers due to the high demand in various fields of industries. In this dissertation, we treat three major pedestrian visions.

First pedestrian detection, detecting pedestrian in the given images, is one of the well-known pedestrian visions. Because many pedestrian visions require the results of pedestrian detection, many researchers have been working to improve it's performance. As a result, pedestrian detection based on deep networks performed well on many challenges. Although there are still difficult issues (miss detection of small pedestrian, heavy occlusion, computational load of deep network), the current pedestrian detection field has achieved significant performance improvements over the past

1. Pedestrian detection



2. Re-identification feature based pedestrian matching

3. Multi-pedestrian tracking

Figure 1.1: Relationship between pedestrian detection, person re-identification and multi-pedestrian tracking.

few years.

Second person re-identification which is intensively studied in recent years is another representative pedestrian vision. In the past, researchers have not achieved satisfactory performance due to the difficult properties of person re-identification (few shot learning, fine-grained image classification). However, with the advent of deep learning, the category of classification achieved significant performance improvements, which significantly affected person re-identification. Improvements in personal re-identification have facilitated research into various pedestrian visions, personal searches and personal re-ranking.

Finally, multi-pedestrian tracking (the pedestrian-focused version of multi-object tracking) is the most classic vision problem in pedestrian vision. Unlike other pedestrian visions, multi-pedestrian tracking is a difficult area to directly improve performance using deep learning. Instead, methods based on the classical tracking-by-detection framework places a great deal on improving tracking performance, it is important to use an accurate pedestrian detector. These three studies are essential and complementary to each other. We are going to solve more complex and difficult tracking problems by combining these three methods, we define this problem as Wide-Area Multi-Pedestrian Tracking (WaMuPeT).

In this dissertation, we propose the tracking scheme for WaMuPeT. To achieve the WaMuPeT, we propose three algorithms for improving the performance of pedestrian localization, person re-identification and multi-pedestrian tracking. First, we propose a novel localization method of pedestrian detection which is more suitable for multi-pedestrian tracking than conventional localization method. Second we propose a novel re-ranking method reflecting ranking order of nearest neighbor set to improve the performance of person re-identification. A good re-identification network generates discriminative features of pedestrians, which help with multi-pedestrian tracking. Finally we propose a novel data association method that can effectively match deep

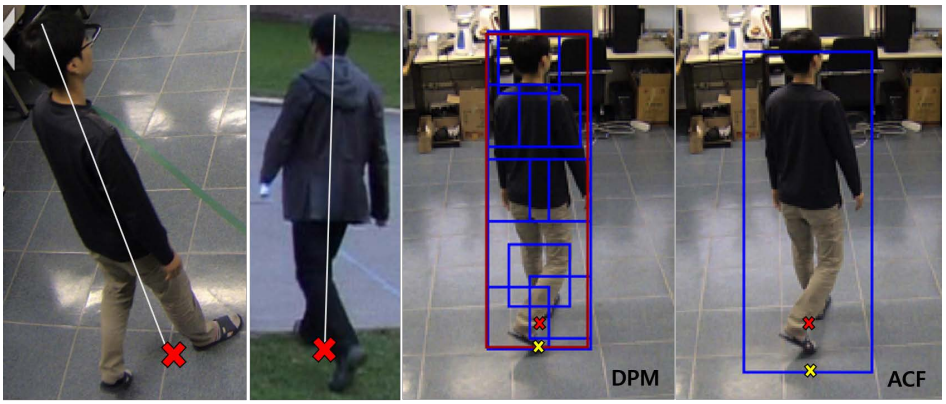


Figure 1.2: Examples of 2-D localization of pedestrian (red 'x' points of left two images) and the points of 2-D localization generated by pedestrian detector (right two images). yellow 'x' points refer to the bottom center points of each pedestrian detectors and the red 'x' points refer to the pedestrian's actual foot coordinates (DPM : Deformable Part Model [1] and ACF : Aggregated Channel Feature [2]).

re-id features for generating hierarchical trajectories.

1.2 Related Works

1.2.1 Localization of Pedestrian Detection

Proper localization of pedestrian detection is the first prerequisite for multi-pedestrian tracking. Depending on the type of detector, the performance of localization is also very various. HOG+SVM [5] is the most classic human detector. It generates the histogram of oriented gradients from image and trains linear support vector machine to classify human and not human. But the detection boxes of HOG+SVM are not quite tight to human body. More advanced version of HOG+SVM are presented by [1] called by DPM (Deformable Part Model). DPM reflects human's body parts into SVM model, so it's bounding boxes are very tight to human body. But it has some computational

load caused by heavy multi-scaling and extracting dense hog-features. In 2014, [2] proposed fast feature pyramid matching method with ACF (Aggregated Channel Feature). By training cascaded boosting trees, ACF detector shows good accuracy and detection speed compared to DPM. With the advent of the deep learning era, pedestrian detection also have entered a new phase. Faster R-CNN [6] is one of the most successful object detector using convolutional neural network. Using ROI (Region Of Interest) pooling, feature map can be effectively refined to scores formed by one hot vector. In general, the recently proposed pedestrian detector generates a tighter bounding box. Note that, all the previous researchers have focused on finding a tight bounding box not on finding the foot points of pedestrian. Actually the foot points of pedestrian are a crucial clue for projecting 2-D position of human to 3-D world coordinates. In our localization work, we focus on this observation and propose a novel localization method.

1.2.2 Pedestrian Feature from Person Re-identification

The Re-ID studies before deep learning have used hand-crafted features [7–10]. With the success of deep learning, recent Re-ID works achieved significant improvements with the help of deep learning networks such as ImageNet [11]. [12] proposed a filter pairing neural network and began to incorporate deep-learning into Re-ID studies. [13] used the siamese network [14] for Re-ID. The siamese network calculates the loss of the image pair using the weight sharing network. [15] proposed a triplet loss. In mini-batch, a triplet, which is composed of an anchor, the positive sample of the anchor and the negative sample of the anchor, is used for calculating the triplet loss. Using this loss, the trained network showed high Re-ID performance. [16] proposed a quadruplet network to further improve the triplet network. In mini-batch, a quadruplet, which is generated by adding a negative sample to a triplet, is used for calculating the quadruplet loss. With this quadruplet loss, the trained network showed higher Re-ID performance than triplet loss. Meanwhile, [17] proposed a batch sampling method

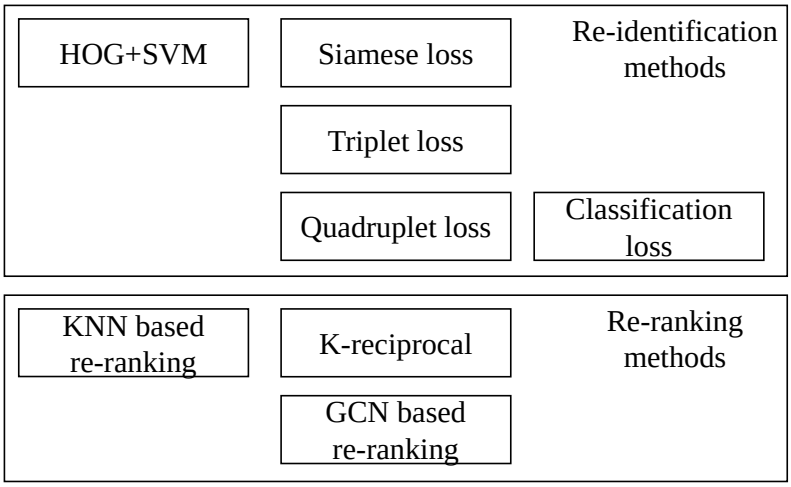


Figure 1.3: Various methods of re-identification and re-ranking methods.

suitable for triplet loss to improve Re-ID performance. [18] introduced a new training loss to classify a person's IDs.f Also, [18] improved the Re-ID performance by training the deep network with negative samples created by a Generative Adversarial Network (GAN) [19]. [20] used GAN for reducing the Re-ID performance loss caused by a bias in the dataset. In [21], they applied a feature attention concept to the network. The network learns which part of the feature should be considered.

Re-ranking methods also affect the generation of discriminative re-id features. [22] proposed a penalty score according to appearance distance for re-ranking. [23] proposed a soft biometric (SB) distance based on semantic information of images for re-ranking. [24] proposed a re-ranking method through optimization based on a discriminant context information analysis. [25] proposed a similarity measure comparing KNNs of a probe and a gallery to improve re-ranking performance. [26] proposed K-reciprocal Nearest Neighbors for further refinement of a KNN-based similarity measure. In [26], they proposed a new *jaccard* distance based on the assumption that probes and galleries with the same ID have many sharing reciprocal neighbors. [27]

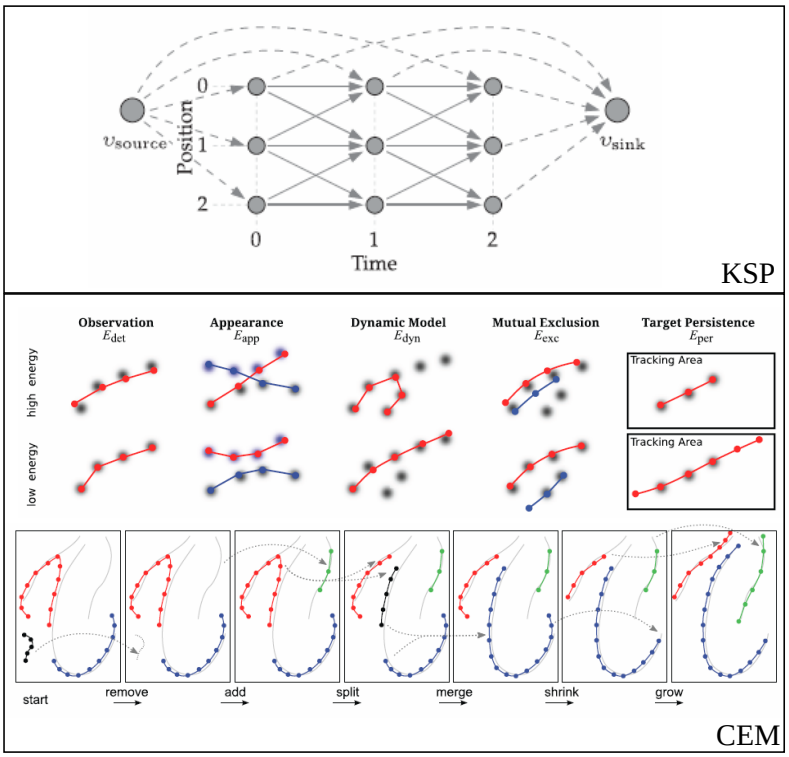


Figure 1.4: Two multi-pedestrian tracking methods targeting single static camera. Top : K-shortest path optimization tracking [3]. Bottom : Continuous energy minimization tracking [4].

proposed an Expanded Cross Neighborhood (ECN)-based re-ranking method, extending a KNN-based similarity measure. Recently, some Re-ID studies target a video images. [28] suggests a network that generates an aggregation image from a video using GAN and solves the classification problem using the aggregation image. [29] suggests a method that extracts a video feature using an attention concept in video.

1.2.3 Multi-Pedestrian Tracking

Multi-pedestrian tracking can be represented by the aggregation of pedestrian detection, person re-identification and data association method Since pedestrian detection and person re-identification are introduced above section, we introduce mainly data association method for multi-pedestrian tracking here. Multi-pedestrian tracking using K-shortest paths (KSP) optimization [3] are proposed in 2011. Generating a probabilistic occupancy map from a detector, KSP reformulates integer programming as a k-shortest node-disjoint paths problem on a directed acyclic graph (DAG). [4] propose a continuous energy minimization (CEM) based on the different components considering observation, appearance, dynamics and some tracking conditions. Compared to KSP, CEM shows relatively better performance because it explicitly defines the crucial factors for multi-pedestrian tracking as an energy function. After achieving good performance in single camera multi-target tracking, many researchers focused on the multi-camera multi-target tracking problem.

1.3 Contributions

In this dissertation, we present various trajectory matching methods for handling camera settings and scene density changes. First, we propose a trajectory matching in overlapping camera settings to find accurate 2-D and 3-D position of pedestrian, which lead the improvement of multi-pedestrian tracking performance. Compared to previous detection method, proposed method find the foot point of pedestrian in a cropped image of bounding box. To implement this, we construct a deep neural network whose output is the 2-D foot point of input pedestrian image. Also the network detect pedestrian while simultaneously localizing 2-D foot point. In our experiments, enhanced 2-D and 3-D localization points of pedestrian leads to the improvement of tracking performance. From the first observation, we can see that it is important to find the 2D

localization point of the pedestrian to get accurate 3D world coordinates and this is the first prerequisite to improve multi-pedestrian tracking.

Second, trajectory matching in overlapping camera settings is proposed to improve the accuracy in person re-identification, which generate deep features that are crucial factor of multi-pedestrian tracking. In the proposed re-ranking method, three key factors contribute to the accuracy improvement. The first factor is the ranking-reflected similarity (RSS) between the ordered set of K-nearest neighbors (OKNN) of a probe and that of a gallery. The second factor is the selection of candidate neighbor sets to construct the OKNNs. The third factor is the re-ranking procedure, where priority is given to the galleries likely to have the same ID as the given probe rather than re-ranking the entire galleries. As validated in the experiments, the three factors in the proposed re-ranking method lead to the improvement of Re-ID accuracy, outperforming the state-of-the-art re-ranking methods. A well-trained re-id network generate a discriminative features to distinguish different pedestrians. This is a good feature for the similarity matrix of multi-pedestrian tracking as our re-ranking method enhances the network's discriminative power and this is second prerequisite to improve multi-pedestrian tracking.

Third, we propose a robust trajectory matching in dense scene settings based on appearance matching. In the process of merging from short-term to long-term trajectory, we propose a appearance based stable boundary selection (SBS) match for merging multiple tracklets into a single trajectory. To achieve best results for SBS match, we design a new cost matrix reflecting the various factor needed for multi-pedestrian tracking (a similarity of deep features, tracking incompatibility, box displacement). Going through multiple stages with SBS and *hungarian* match, the proposed method generate robust trajectories. In our experiments, the proposed method demonstrates state-of-the-art tracking performance in public datasets.

1.4 Thesis Organization

In Chapter 2, as for the problem statement, we represent the challenges of pedestrian detection in Section 2.1, re-ranking of person re-identification in Section 2.2 and multi-pedestrian tracking in Section 2.3. The proposed 2-D and 3-D localization method of pedestrian detection is represented in Chapter 3. In Chapter 4, we explain the ranking reflected re-ranking method to enhance person re-identification. Using enhanced localization and re-identification results, we introduce the method for improvement the multi-pedestrian tracking performance in Chapter 5. In addition, we explain the proposed multi-pedestrian tracking framework using stable boundary selection. In Chapter 6, we show the experimental results including qualitative and quantitative results through public datasets. Conclusions and future works of this dissertation are presented in Chapter 7.

Chapter 2

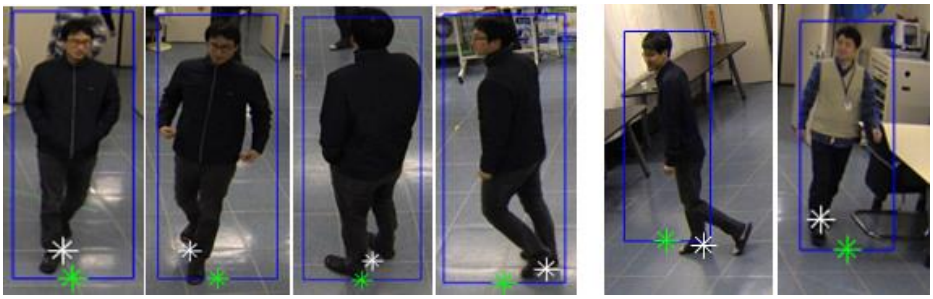
Problem Statements

2.1 Trajectory Matching in Overlapping Camera Settings

2.1.1 Challenges

In overlapping Camera Settings, tracking-by-detection framework for multi-camera surveillance has been successfully achieved in recent years. In the tracking-by-detection framework, pedestrian detection and multi-camera multi-target tracking (MCMTT) are combined to find and track pedestrians. In MCMTT, 3-D localization is essential to associate detections, because most MCMTT methods [3, 30–36] try to solve the problem in 3-D space. Nevertheless, most studies overlook the error in estimating the 3-D position of an object precisely.

In general, the 3-D position is estimated by the projection of a 2-D pedestrian ground position (2-D PGP) into 3-D space using a camera matrix. In the projection, bottom center points of 2-D detections are usually used as a 2-D PGP estimate. However, this estimate frequently causes 3-D localization errors. Figure 2. 1(a) shows the results of the conventional pedestrian detectors (LDCF [37]). In Figure 2. 1(a), the 2-D



(a) Pose variation (b) Image distortion

Figure 2.1: Left : Compared to the actual pedestrian 2-D ground position (white), the one given by the conventional detections (green) varies depending on the posture. Right : Image distortion misleads the estimation of the pedestrian 2-D ground position.

PGP varies depending on the pedestrian's posture. This means that it is difficult to get a precise 2-D PGP without considering the posture of the person in the bounding box. In addition, distortion in the image is caused by camera lens refraction or change in angle of view. Figure 2. 1(b) shows that the 2-D PGP can be changed by an image distortion. From the above observations, we can infer that it is necessary to find 2-D PGP considering the appearance in the bounding box.

For most MCMTT methods, the tracking input is a set of bounding boxes given by a pedestrian detector. However, the detector may generate a wrong 2-D PGP due to the issues mentioned above. Then most existing MCMTT methods have no chance to correct the 3-D localization error caused by the wrong 2-D PGPs. 3-D localization accuracy has been improved by smoothing the tracking result as a post-processing [30–32] function. However, the method does not solve a fundamental problem, causing localization errors. To decrease localization error and improve tracking performance, it is required to estimate an accurate 2-D PGP.

2.1.2 Approach for the challenges

Inspired by this investigation, we tackle the challenges using a deep-learning network that performs 2-D localization as well as the pedestrian detection. To this end, we try to design a deep-learning network which includes a detection network and a localization network. The detection network is designed to yield a detection score of a pedestrian from the proposals given by the object proposal method. And the localization network is designed to estimate the 2-D PGP from the pedestrian's posture in the input bounding box. Through the deep networks, we aims to achieve accurate 3-D localization and tracking from the accurate 2-D PGPs produced by the deep network.

2.2 Trajectory Matching in Non-Overlapping Camera Settings

2.2.1 Challenges

In Non-Overlapping Camera Settings, person re-identification (Re-ID) is an important and challenging problem in many computer vision applications such as sports analysis, surveillance systems, and social interaction analysis. The purpose of the Re-ID is to obtain the best match between the probe set having known identities and the gallery set having unknown identities. In recent years, with the advance of deep learning, Re-ID research has made a lot of progress [11–13, 15–18, 20, 21, 38, 39]. Deep-learning-based Re-ID algorithms have improved performance remarkably over the existing methods, nevertheless, some challenging problems remain. One of the challenging problems is the *appearance ambiguity problem*. This problem arises when a hard negative gallery, which is very similar to the probe but has a different ID from the given probe, is wrongly matched with the probe. The appearance ambiguity problem results from a lack of training data per an ID in most Re-ID dataset.

Figure 2.2: Examples of the appearance ambiguity problem. Each color boxes means probe images (blue), true gallery images (green) and false gallery images (red). As shown in above figure, there are false matching that even a person might mistake at first glance caused by same color of clothes or accessories.

2.2.2 Approach for the challenges

To solve the appearance ambiguity problem, many re-ranking methods have been proposed as a post-processing of Re-ID methods such as [22–27, 40, 41]. Adding new clues (e.g., neighbor information of a probe or gallery) based on a feature distance, re-ranking modifies the initial ranking list to improve the ranking accuracy of Re-ID.

Recent re-ranking methods are based on the distance between K-nearest neighbors (KNN) of a probe and those of a gallery [14], which is motivated by the assumption that the two KNNs share many neighbors if the probe and gallery have the same identity [26]. Distance between two KNNs helps to resolve the ambiguity of appearance compared to the distance between two appearance features generated by the deep network. This is because the KNN set contains similar deformed images of the probe or gallery. The deformed images provide good sources for re-ranking because they provide various poses sharing similar features (same clothes, accessories, bags, etc.). However, the appearance ambiguity problem has not been completely solved by the re-ranking based on the distance between two KNNs, where the distance depends on the number of shared elements. Some hard negative galleries (high-ranked neighbors having a different ID from the given probe's ID) may have KNNs similar to those of the probe, so they can not be easily distinguished from positive galleries. To distinguish hard negative galleries to positive galleries, more precise metric is needed. According to our empirical observation, the ordering of the shared neighbors in KNN takes an important role to distinguish the hard cases in contrast to the existing distance metric depending on only the number of shared neighbors between two KNNs. That is, a shared neighbor tends to have the same rank in two KNNs when the gallery is a true positive. To resolve the ambiguity problem,we attempt to complement the KNN-based method by following the observation.

To this end, we try to find a new metric to measure the similarity between the ordered set of K-nearest neighbors (OKNN) of the probe and the OKNN of the gallery. Instead of using KNN, we use OKNN to use ranking of KNN for measuring similarity between two OKNNs. Proposed similarity measure is calculated by a weighted rank-sum of shared neighbors between two OKNNs. By calculating a weighted rank-sum of shared neighbors, we can estimate correlation of neighbor's order between two OKNNs. What the weighted rank-sum means here is that the shared neighbor in

the front order derives larger similarity than that in the rear order. By adding criteria that OKNN should be similar in order, we design a finer metric for similarity. In the re-ranking procedure, based on the proposed similarity between two OKNN sets, we rearrange the initial ranking list obtained by the Re-ID network. Additionally, we give priority to the galleries likely to be a true positive rather than the entire galleries by calculating *first responsible gallery set*. In the experiment, we show that the proposed re-ranking method improves performance by adding the re-ranking procedure to the state-of-the-art methods.

2.3 Robust Trajectory Matching in Dense Scene Settings

2.3.1 Challenges

With a single static camera, many previous researchers try to solve the multi-pedestrian tracking problem. Given detection boxes from start to end image frames of a dataset or a sequence, the goal of multi-pedestrian tacking is to identify all the pedestrians and track them by linking temporal association of the detection boxes. The main challenge of multi-pedestrian tracking is to solve the occlusion problem which causes tracking error like ID switch or trajectory fragment. Since detection bounding boxes are not perfectly given, most of multi-pedestrian tracking algorithms try to solve the occlusion problem by using the acceleration information of the pedestrian. And it is also a valid approach to use the pedestrian matching in adjacent images for solving the occlusion problem.

Previous researchers have proposed various methods to overcome above challenges. One of most successful multi-pedestrian tracking method is a energy minimization based algorithm. A. Milan et al. [4] have proposed continuous energy minimization (CEM) for multi-pedestrian tracking. To effectively handle the collision between pedestrians, CEM introduce the various energy functions representing different

Figure 2.3: Three cases of multi-pedestrian tracking error caused by occlusion.

collision situations (energy functions based on appearance, dynamics, mutual exclusion, target persistence). They achieve reasonable tracking performance in PETS2009-S2L1, S2L2, and TUD-Stadtmitte dataset, but some tracking issues occurring ID switch are remain.

ID switch, change ID between adjacent targets, is one of the main challenges of multi-pedestrian tracking. The main reason why the previous tracking algorithms have failed to overcome the ID-switch problem is that the methods do not use deep features effectively. With a rising of the performance of deep re-identification, deep features of pedestrians are very discriminative enough to use as a main factor of tracking algorithm. Surely there are many multi-pedestrian tracking algorithms treating deep features as a appearance model. But we observe that the discriminative power of pedes-

trian's deep features are maximized when pedestrians are grouping. When comparing two small grouped pedestrians, the similarity matrix can be generated by representing one pedestrian group as a row and another pedestrian as a column. In a few adjacent frames, the pedestrian appearance is not change drastically so features in one pedestrian are similar. It makes a solid comparison between two short-term trajectories.

2.3.2 Approach for the challenges

In this dissertation, we propose a novel multi-pedestrian tracking framework based on a tracklet (very short trajectory) matching. Our framework is based on the divide and conquer algorithm. Since it is not easy to find whole trajectory of a pedestrian at once, we divide whole image sequences into many small sequences to generate short-term trajectories. Through small sequences, since the pedestrian appearance is not change drastically, appearance based matching can be performed robustly, which generates robust short-term trajectories. Here, We propose a new appearance matching method that reflects the characteristics of one pedestrian group with similar deep features. We define this matching method as a stable boundary selection (SBS). Based on the robust short-term trajectories, we generate mid-term trajectories in mid-sized sequences. Finally we can generate long-term trajectories by matching all the mid-term trajectories in consequence image windows. The strong point of this divide and conquer framework is that the robust short-term trajectories yield to generate more robust longer trajectories. Compared to one to one feature matching, trajectory to trajectory feature matching guarantees accurate matching results. This is because that a trajectory containing missed detections or detections of various pose is complemented by other intact detections. So when conducting trajectory to trajectory matching, we can reduce the matching errors caused by a few missed detection and a feature variation. We also consider the restoration of missed detections and the smoothness of trajectories for final long-term trajectories to enhance tracking accuracy.

Chapter 3

Trajectory Matching in Overlapping Camera Settings

3.1 Overall Scheme

To enhance the trajectory matching performance at 3-D space in overlapping camera settings, we propose a novel deep-learning network that performs 2-D localization as well as the pedestrian detection. As shown in Figure 3.1, the proposed deep-learning network (DL-net) is designed to be composed of two networks: a detection network (D-net) and a localization network (L-net). D-net yields a detection score of a pedestrian from the proposals given by the object proposal method. L-net estimates the 2-D PGP from the pedestrian's posture in the input bounding box. In addition, an attentional pass filter (APF) is introduced to pass a detection candidate that may be a pedestrian. Connecting D-net to L-net, APF can reduce the load to compute feature of a detection candidate that might not be a pedestrian. From the output of DL-net, the multi-camera multi-target tracking (MCMTT) problem is developed through the min-cost network flow approach followed by [32]. Accurate 3-D localization and tracking

can be achieved from the accurate 2-D PGPs given by the proposed network. The improvement of 3-D localization and tracking is verified through the experiments by showing localization error and public tracking performance measures.

3.2 Network Design

To detect pedestrians and perform localization at the same time, we propose two loss functions to train the proposed network: detection loss function and localization loss function. Detection loss learns whether an input candidate box is a pedestrian or not. Localization loss learns grounding positions which indicates the point where the pedestrian supports the ground. We observe that the box regression loss in previous studies [42] has a limitation to learn a well-localized point. The proposed localization loss can learn an accurate grounding positions from training data. The attentional pass filter (APF) is introduced to connect detection network (D-net) and localization network (L-net) effectively. APF delivers only the features related with the pedestrian presumed by D-net to L-net. As a result, D-net and L-net are efficiently connected through APF without computational redundancy.

Detection Network (D-net). A training set is defined by $\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, ...\}$, where $\mathbf{s}_i = \{I_i, c_i, t_i, b_i\}$. I_i refers to an input image, c_i the camera, and t_i the time index. $b_i = (x_i, y_i, w_i, h_i)$ is a detection box with the left top coordinates (x_i, y_i), width w_i and height h_i. A training sample $\mathbf{s}_i$ is fed to VGG-16 network [43] to generate feature map X_i from I_i. To crop specific box region from F_i, ROI-pooling layer takes the input as F_i and b_i. An output of ROI-pooling layer [42] is $\mathbf{f}_i$ which means pooled ROI feature vector from F_i corresponding to b_i. $\mathbf{f}_i$ is fed to D-net to generate a discrete probability p_i which is the output of D-net. p_i is a 2-dimensional vector, where non-pedestrian probability $p_{i,0}$ and pedestrian probability $p_{i,1}$.

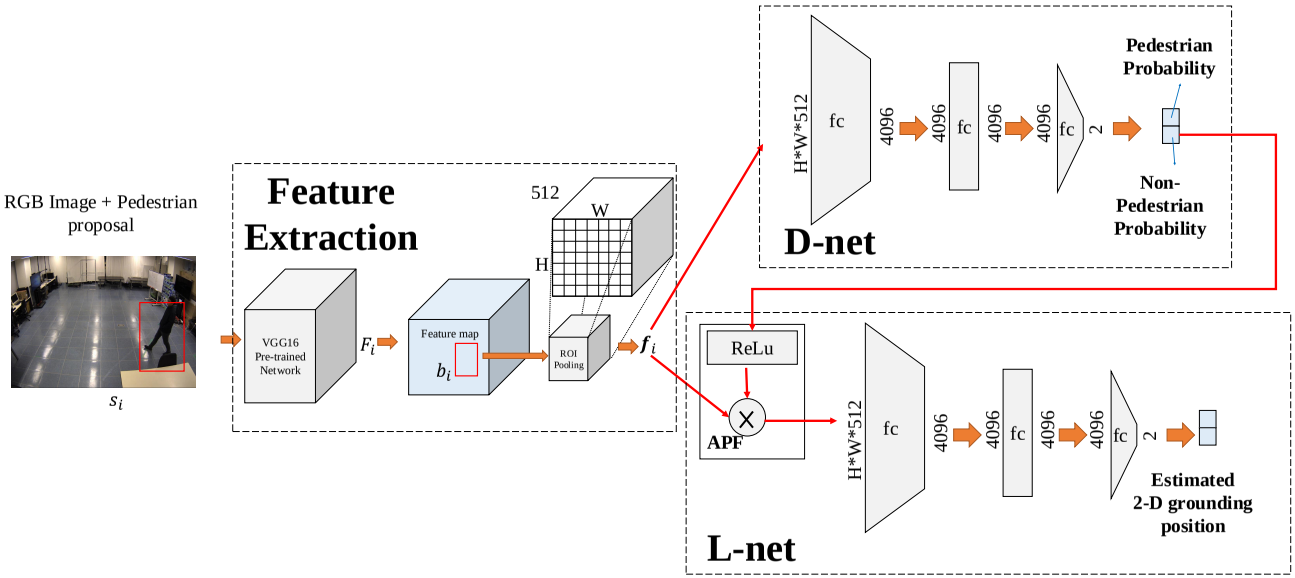


Figure 3.1: Architecture of deep-learning network (DL-net) composed of detection and localization network. Input image and pedestrian proposals are fed to the input of DL-net. D-net yields detection scores and L-net yields a grounding position from the feature of input proposal. Attentional pass filter (APF) only delivers proposals that are likely to be pedestrians.

Specifically, D-net has the following structure: FC(H $\times$ W $\times$ 512, 4096)-ReLu-DR-FC(4096, 4096)-DR-ReLu-FC(4096, 2), followed by fast rcnn fully-connected network [42]. FC means a fully connected layer and DR means a drop out layer. Compared with [42], D-net does not perform box regression. Instead of using box regressor, L-net is introduced for an accurate localization.

Localization Network (L-net). Unlike D-net, L-net performs regression task to estimate points indicating 2-D PGP. In the first step of L-net, p_i and $\mathbf{f}_i$ are fed to APF defined by

$$APF(p_i, \mathbf{f}_i) = ReLu(p_{i,1})\mathbf{f}_i. \tag{3.1}$$

The output of APF is a feature map determined as a pedestrian by D-net. It is inefficient for all inputs to be passed through L-net because most of the detection inputs are not a pedestrian class. APF has a role to prevent low confidence $\mathbf{f}_i$ to be fed to the fully-connected layer. The output of L-net is a 2-dimension vector $\tilde{g}_i = (\tilde{g}_i^x, \tilde{g}_i^y)$. $\tilde{g}_i$ indicates the relative coordinate of a 2-D PGP. And the absolute coordinate g_i is given by $g_i = (\tilde{g}_i^x + x_i, \tilde{g}_i^y + y_i)$. In section 3.3, we use g_i as an i-th observation of tracking inputs. Compared to box regressor, L-net estimates the 2-D PGP instead of tightening the bounding box. L-net is configured similar with D-net by following structure: APF-FC(H $\times$ W $\times$ 512, 4096)-ReLu-DR-FC(4096, 4096)-DR-ReLu-FC(4096, 2).

3.3 MCMTT with Proposed Network

Track Assignment Formulation. MCMTT in this paper is formulated as an optimization problem to assign the detections to multiple tracks corresponding to pedestrians. Assume that N_c overlapped cameras are set at indoor space. Then N_c images are generated every frame. Let g_i be the 2-D PGP of the detection b_i. Here let $\mathcal{U} = \{\mathbf{u}_1, \mathbf{u}_2, ...\}$ be the given information for the tracking formulation, where $\mathbf{u}_i = \{b_i, g_i, c_i, t_i\}$, c_i is the camera index and t_i is the time index. With $\mathbf{u}_i$, we define the k-th association set

$\mathcal{D}_k$ by

$$\mathcal{D}_k := \{\mathbf{u}_i | \forall \mathbf{u}_i, \mathbf{u}_j \in \mathcal{U}, i \neq j : c_i \neq c_j \wedge t_i = t_j\}, \tag{3.2}$$

which means a set of 2-D detections at the same time, but in different view, which can be a candidate set of detections from different cameras for a pedestrian. MCMTT is a problem to find a linked set of $\mathcal{D}_k$ associated with each target. In this paper, the tracking problem is formulated by min-cost flow problem similar to [32].

$$\mathcal{F}^* = \arg\min_{\mathcal{F}} \sum_k C_k f_k + \sum_k C_{en,k} f_{en,k} + \sum_{k,l} C_{k,l} f_{k,l} + \sum_k C_{ex,k} f_{ex,k}. \tag{3.3}$$

Equation (4.3) implies that the association set $\mathcal{D}_k$ has flow f_k with the cost C_k. $f_{k,l}$ and $C_{k,l}$ are the flow and cost of the temporal edge between $\mathcal{D}_k$ and $\mathcal{D}_l$. $f_{en,k}, C_{en,k}$ and $f_{ex,k}, C_{ex,k}$ are the flow and cost of the source and sink respectively. f is a binary integer value, where $f = 1$ implies that the corresponding edge is a part of the corresponding trajectory and $f = 0$ implies that the edge is not used.

Cost Design. We design new C_k, which means the reconstruction cost of association set $\mathcal{D}_k$.

$$C_k = C(\mathcal{D}_k) = \lambda_{rec} \frac{\sum_{u_i \in \mathcal{D}_k} dist(g_i, G_k; c_i)}{V_k}, \tag{3.4}$$

where λ_{rec} is a weighting constant. $dist(g_i, G_k; c_i)$ is defined by distance between a 3-D point G_k and the line back-projected from g_i using camera calibration information c_i. V_k is the number of cameras where G_k is visible. G_k is the 3-D vector minimizing the distance between the lines generated from all g_i in $\mathcal{D}_k$. That is,

$$G_k = \arg\min_{G} \sum_{u_i \in \mathcal{D}_k} dist(g_i, G; c_i). \tag{3.5}$$

G_k means the virtual 3-D PGP of $\mathcal{D}_k$. $dist(g_i, G_k; c_i)$ is given by

$$\begin{aligned} &dist(g_i, G_k; c_i) = \\ &\frac{||(G_k - \Phi^{c_i}(g_i, z_{min})) \times (\Phi^{c_i}(g_i, z_{max}) - \Phi^{c_i}(g_i, z_{min}))||_2}{||\Phi^{c_i}(g_i, z_{max}) - \Phi^{c_i}(g_i, z_{min})||_2}, \end{aligned} \tag{3.6}$$

where $\Phi^c(g, z)$ is a projection function related to camera c which deliver from the 2-D coordinate g to 3-D coordinate $(\cdot, \cdot, z)$. z_{min} and z_{max} are the constants of the minimum and maximum height of a pedestrian. $dist(g_i, G_k; c_i)$ implies the error between G_k and g_i in 3-D space . The reconstruction error $C(\mathcal{D}_k)$ means the average error of 3-D reconstruction from each cameras. $C(\mathcal{D}_k)$ mainly depends on how accurate g_i is. Most of MCMTT methods, g_i is obtained from b_i by the bottom center of the detection box, i.e.,

$$g_i = (x_i + \frac{w_i}{2}, y_i + h_i). \tag{3.7}$$

In our works, instead of using Equation (4.7), L-net yields g_i as

$$g_i = \text{L-Net}(p_i, \mathbf{f}_i). \tag{3.8}$$

While Equation (4.7) is a linear mapping from bounding box to 2-D PGP, Equation (4.8) is non-linear mapping from an image to 2-D PGP. The fact that using image-level feature and non-linear mapping function give a better result than using (4.7).

Equation (4.3) is a binary integer programming problem (BIP) like in [32]. We use branch-and-cut procedure to solve Equation (4.3), which is implemented in the Gurobi optimization library [44].

Chapter 4

Trajectory Matching in Non-overlapping Camera Settings

4.1 Overall Scheme

In Fig. 4.2, the overall scheme of the proposed method is depicted to state the problem in this paper. The purpose of Re-ID is to assign an identity to a person in the image set of unidentified persons (probes) by referring to the known identities from the image set of identified persons (galleries). First, all images of the probes and galleries are used as the inputs of a deep neural network for feature extraction. Using the appearance features of probes and galleries obtained through the network, we form a probe set and a gallery set, defined by

$$\begin{aligned} \mathcal{P} &= \{\, p_i \mid p_i = \{\, x_{p_i}, l_{p_i}, c_{p_i}, \,\}, i = 1, 2, ..., N_{\mathcal{P}} \,\}, \\ \mathcal{G} &= \{\, g_i \mid g_i = \{\, x_{g_i}, l_{g_i}, c_{g_i}, \,\}, i = 1, 2, ..., N_{\mathcal{G}} \,\}, \end{aligned} \tag{4.1}$$

where x represents the appearance feature, l represents the label and c represents the camera index of each image in the probe or gallery set, whereas $N_{\mathcal{P}}$ and $N_{\mathcal{G}}$ are the number of elements in probe and gallery set, respectively.

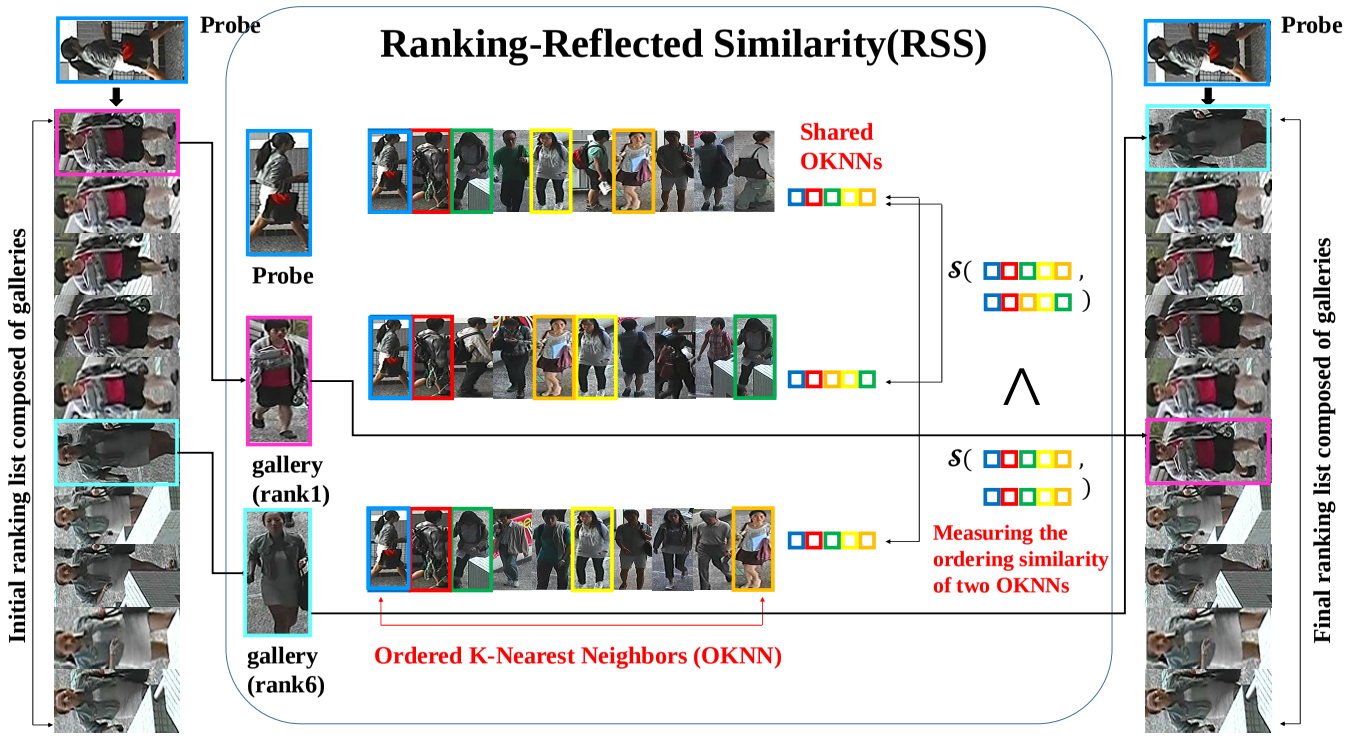


Figure 4.1: Illustration of re-ranking with the ranking-reflected similarity. Comparing the images in three OKNNs, the shard neighbor (green box) between OKNN of probe and that of 6-th ranked gallery has higher ranking than that (yellow box) of 1-th ranked gallery in the initial ranking list. Hence according to the proposed ranking-reflected similarity, their rankings are reversed in the final ranking list.

In the Re-ID problem, the cameras used in the probe set and gallery set are set dis-jointly, satisfying $c_{p_i} \neq c_{g_j}$. Then we calculate the appearance distance for all combination pairs of appearance features. The $(i, j)-$th element of the distance matrix between $\mathcal{A}$ and $\mathcal{B}$ ($\mathcal{D}^{\mathcal{A},\mathcal{B}}$) is defined by

$$\mathcal{D}_{i,j}^{\mathcal{A},\mathcal{B}} = d(x_{a_i}, x_{b_j}), a_i \in \mathcal{A}, b_j \in \mathcal{B}, \tag{4.2}$$

where the appearance distance is generally measured by the *mahalanobis* distance as

$$d(a_i, b_j) = \sqrt{(x_{a_i} - x_{b_j})^T M (x_{a_i} - x_{b_j})}, \tag{4.3}$$

where M is a positive semi-definite matrix. When $M = I$, the distance becomes Euclidean distance. We use the Euclidean distance in all the experiments.

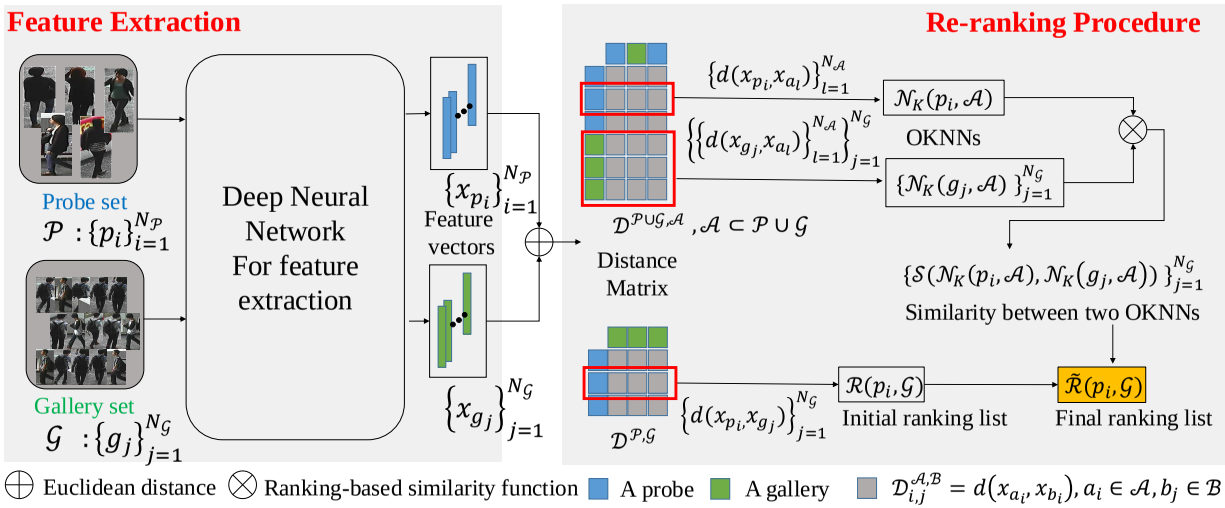


Figure 4.2: Overall scheme of our re-ranking procedure. For feature extraction, we employ ResNet-50 as the baseline network. After extracting the features of probes and galleries, we generate OKNNs of probes and galleries. We use the generated OKNNs to calculate the proposed ranking-based similarity metric and obtain the final ranking list. The key factors that can improve performance in our method are as follows: the selection of the candidate neighbor set $\mathcal{A}$, the ranking-reflected similarity metric, and the re-ranking procedure that priority is given to the galleries likely to have the true ID for the given prove.

From the distance matrix $\mathcal{D}^{\mathcal{P},\mathcal{G}}$, we obtain an initial ranking list $\mathcal{R}(p_i, \mathcal{G})$ of gallery set $\mathcal{G}$ for a sample $p_i \in \mathcal{P}$, which is defined by

$$\mathcal{R}(p_i, \mathcal{G}) = \{\, g_1^{p_i}, g_2^{p_i}, ..., g_{N_\mathcal{G}}^{p_i} \,\}, \tag{4.4}$$

where $d(p_i, g_k^{p_i}) \leq d(p_i, g_{k+1}^{p_i})$ for $g_k^{p_i} \in \mathcal{G}$.

If the initial ranking list is perfect, we can find true labels for all galleries. However, since the Re-ID problem is a few-shot learning problem that lacks learning data, the learning of the feature extraction network is incomplete and the feature is incomplete to obtain the perfect ranking list. One of the failure cases of the initial ranking list is the appearance ambiguity problem. This problem occurs when two images having similar appearance have different IDs from each other.

To overcome the appearance ambiguity problem, re-ranking methods have been proposed, where new appearance distance metrics were designed to modify the initial ranking list [22–24, 40, 41]. Recently, instead of appearance distance, a set-distance between neighbor sets of probe p_i and gallery g_j, was introduced and this greatly contributed to the improvement of Re-ID performance [26, 27]. They have designed the set-distance to be small when the number of shared neighbors of two neighbor sets is large. However, this method may not work for a hard negative gallery whose neighbor set is similar to that of the given probe.

To handle this problem, this paper proposes a new similarity metric that measures the similarity between two ordered sets of K-nearest neighbors (OKNN) of a probe and a gallery. To this end, from $\mathcal{D}^{\mathcal{P}\cup\mathcal{G},\mathcal{A}}$, $\mathcal{A} \subset \mathcal{P}\cup\mathcal{G}$, we generate OKNNs, $\mathcal{N}_K(p_i, \mathcal{A})$ and $\{\mathcal{N}_K(g_j, \mathcal{A})\}_{j=1}^{\mathcal{G}}$, where $\mathcal{A}$ refers to a candidate set of neighbors to find neighbors. Through the proposed measure $\mathcal{S}(\mathcal{R}_K(p_i, \mathcal{A}), \mathcal{R}_K(g_j, \mathcal{A}))$ in (4.8), we measure not only how many neighbors are shared by two OKNNs, but also how similar their ranking is. Based on the calculated similarity, the final ranking list $\tilde{\mathcal{R}}(p_i, \mathcal{G})$ is obtained by re-ranking the initial ranking list $\mathcal{R}(p_i, \mathcal{G})$.

4.2 Proposed Method

4.2.1 Proposed Similarity Metric

The ordered neighbor set of $p_i \in \mathcal{P}$ or $g_j \in \mathcal{G}$, which obtained from the candidate neighbor set $\mathcal{A} \subset \mathcal{P} \cup \mathcal{G}$, is defined by

$$\begin{aligned} \mathcal{N}(p_i, \mathcal{A}) &= \{\, a_1^{p_i}, a_2^{p_i}, ..., a_{N_\mathcal{A}}^{p_i} \,\}, \\ \mathcal{N}(g_j, \mathcal{A}) &= \{\, a_1^{g_j}, a_2^{g_j}, ..., a_{N_\mathcal{A}}^{g_j} \,\}, \end{aligned} \tag{4.5}$$

where $d(p_i, a_k^{p_i}) \leq d(p_i, a_{k+1}^{p_i})$, $d(g_j, a_k^{g_j}) \leq d(g_j, a_{k+1}^{g_j})$ for $a_k^{p_i}$, $a_k^{g_j} \in \mathcal{A}$. From the ordered neighbor sets $\mathcal{N}(p_i, \mathcal{A})$ and $\mathcal{N}(g_j, \mathcal{A})$, the ordered K-nearest neighbor set (OKNN) of p_i and that of g_j are obtained as

$$\begin{aligned} \mathcal{N}_K(p_i, \mathcal{A}) &= \{\, a_k^{p_i} \mid a_k^{p_i} \in \mathcal{N}(p_i, \mathcal{A}), k = 1, ..., K \,\}, \\ \mathcal{N}_K(g_j, \mathcal{A}) &= \{\, a_k^{g_j} \mid a_k^{g_j} \in \mathcal{N}(g_j, \mathcal{A}), k = 1, ..., K \,\}, \end{aligned} \tag{4.6}$$

Now we propose a new similarity measure between the two OKNNs, i.e., $\mathcal{N}_K(p_i, \mathcal{A})$ and $\mathcal{N}_K(g_j, \mathcal{A})$. The core idea to develop the similarity measure is based on the following conjectures.

Conjecture 1: *If g_j has the same ID as p_i, the number of shared neighbors (neighbors appearing in both OKNNs) between the two OKNNs is larger than that for the other gallery $g_{l \neq j}$.*

Conjecture 2: *If g_j has the same ID as p_i, the ranking in OKNN of g_j is similar to that of p_i and the shared neighbors are arranged in the same order.*

Based on these conjectures, we develop a similarity between two OKNNs. From conjecture 1, the similarity between the two OKNN sets is designed to be proportional to the number of shared neighbors between two OKNN sets. And from conjecture 2, the similarity is defined to be proportional to the ranking of the shared neighbors between the two OKNN sets.

First, to count the shared neighbors between $\mathcal{N}_K(p_i, \mathcal{A})$ and $\mathcal{N}_K(g_j, \mathcal{A})$, we define a matching function (M-function) which indicates whether or not $a_k^{p_i} \in \mathcal{N}_K(p_i, \mathcal{A})$ belongs to $\mathcal{N}_L(g_j, \mathcal{A}), L \leq K$. The M-function is given by

$$\mathcal{M}(a_k^{p_i}, \mathcal{N}_L(g_j, \mathcal{A})) = \begin{cases} 1, & \text{if } a_k^{p_i} \in \mathcal{N}_L(g_j, \mathcal{A}) \\ 0, & \text{otherwise.} \end{cases} \tag{4.7}$$

By summing the M-function in (4.7) for all $k \in [1, K]$, we can count the number of shared neighbors in the two OKNNs, which can be a similarity metric satisfying conjecture 1. To meet conjecture 2, we propose a new similarity metric reflecting the ranking order based on the cumulative summing of the M-function for all $\mathcal{N}_L(g_j, \mathcal{A})$, $L = [1, K]$. This similarity metric is referred to as ranking-reflected similarity (RRS) and is expressed by

$$\mathcal{S}(\mathcal{N}_K(p_i, \mathcal{A}), \mathcal{N}_L(g_j, \mathcal{A})) = \sum_{k=1}^{K} \sum_{L=k}^{K} \mathcal{M}(a_k^{p_i}, \mathcal{N}_L(g_j, \mathcal{A})). \tag{4.8}$$

By this RRS, the more similar the ranking of a OKNN is to that of the other OKNN, the larger is the similarity between the two OKNNs. In addition, a shared neighbor in the front order is weighted by counting it several times.

4.2.2 Selection of $\mathcal{A}$

The configuration of the OKNN set varies depending on how the candidate neighbor set $\mathcal{A}$ is constructed. The choice of $\mathcal{A}$ affects the performance of re-ranking because the similarity measure depends on $\mathcal{A}$. In Re-ID problem, we can make three configuration of $\mathcal{A}$: $\mathcal{P}$, $\mathcal{G}$ and $\mathcal{P} \cup \mathcal{G}$. We have adopted the candidate set $\mathcal{P} \cup \mathcal{G}$, which is an experimentally best combination of the three cases. The discussions on this ablation experiment are presented in Sec. 4.3.

4.2.3 Re-ranking Procedure

In this subsection, we explain how re-ranking is performed with the proposed similarity metric. The goal of re-ranking is to find $\tilde{\mathcal{R}}(p_i, \mathcal{G})$, using a new metric that shows better Re-ID performance than $\mathcal{R}(p_i, \mathcal{G})$. The direct approach is to calculate the similarity between the OKNN set of p_i and OKNN sets of all g_j in $\mathcal{G}$, and then obtain $\tilde{\mathcal{R}}(p_i, \mathcal{G})$ sorted in order of large similarities. However, we claim that it should be more accurate to give priority in re-ranking to the galleries likely to be true for a given probe rather than the entire galleries.

To choose the galleries likely to be true from the entire gallery set, a first-responsible gallery set w.r.t the probe p_i is defined by

$$\mathcal{G}_1^{p_i} = \{\, g_j \mid p_i = p_1^{g_j}, p_1^{g_j} \in \mathcal{N}(g_j, \mathcal{P}) \,\}\,. \tag{4.9}$$

The remaining gallery set w.r.t the probe p_i is defined by

$$\mathcal{G}_r^{p_i} = \mathcal{G} - \mathcal{G}_1^{p_i}. \tag{4.10}$$

The first responsible gallery set w.r.t p_i contains g_js whose OKNN includes p_i as the neighbor in the first order. Therefore, the gallery that belongs to the first responsible gallery set w.r.t p_i is more likely to have the same ID as p_i than the remaining galleries. Motivated from the above discussion, we divide the given gallery set $\mathcal{G}$ into $\mathcal{G}_1^{p_i}$ and $\mathcal{G}_r^{p_i}$. Then we generate the re-ranking lists $\tilde{\mathcal{R}}(p_i, \mathcal{G}_1^{p_i})$ and $\tilde{\mathcal{R}}(p_i, \mathcal{G}_r^{p_i})$ through the proposed similarity metric in (4.8). The final ranking list $\tilde{\mathcal{R}}(p_i, \mathcal{G})$ is determined by sequentially listing the two ranking lists. In **Algorithm 1**, the pseudo code for the proposed re-ranking procedure is given.

Algorithm 1: Re-ranking using ranking-reflected similarity metric

input : $\mathcal{P}, \mathcal{G}$

output: $\tilde{\mathcal{R}}(p_i, \mathcal{G}), \forall p_i \in \mathcal{P}$

selection of $\mathcal{A} = \mathcal{P}$;

generate $\mathcal{N}_K(p_i, \mathcal{A}), \mathcal{N}_K(g_j, \mathcal{A})$;

for $\forall p_i \in \mathcal{P}$ **do**

  Divide $\mathcal{G}$ into $\mathcal{G}_1^{p_i}$ and $\mathcal{G}_r^{p_i}$ from (4.9);

  for $\forall g_j \in \mathcal{G}_1^{p_i}$ **do**

    calculate $\mathcal{S}(\mathcal{N}_K(p_i, \mathcal{A}), \mathcal{N}_K(g_j, \mathcal{A}))$ from (4.8);

  end

  calculate $\tilde{\mathcal{R}}(p_i, \mathcal{G}_1^{p_i})$ according to $\{\mathcal{S}(\mathcal{N}_K(p_i, \mathcal{A}), \mathcal{N}_K(g_j, \mathcal{A}))\}_{j=1}^{N_{\mathcal{G}_1^{p_i}}}$;

  for $g_j \in \mathcal{G}_r^{p_i}$ **do**

    calculate $\mathcal{S}(\mathcal{N}_K(p_i, \mathcal{A}), \mathcal{N}_K(g_j, \mathcal{A}))$ from (4.8);

  end

  calculate $\tilde{\mathcal{R}}(p_i, \mathcal{G}_r^{p_i})$ according to $\{\mathcal{S}(\mathcal{N}_K(p_i, \mathcal{A}), \mathcal{N}_K(g_j, \mathcal{A}))\}_{j=1}^{N_{\mathcal{G}_r^{p_i}}}$;

  $\tilde{\mathcal{R}}(p_i, \mathcal{G}) = \{\tilde{\mathcal{R}}(p_i, \mathcal{G}_1^{p_i}), \tilde{\mathcal{R}}(p_i, \mathcal{G}_r^{p_i})\}$;

end

Chapter 5

Robust Trajectory Matching in Dense Scene Settings

5.1 Overall Scheme

First we define the notations. D denotes the set of all detections in all frames. Each detection $d_i \in \mathbf{D}$ is defined by a vector $d_i = (x_i, y_i, w_i, h_i, s_i, f_i)$ where x_i, y_i, w_i, h_i represent the left top coordinates, width and height of a bounding box and s_i, f_i denote the detection and frame index respectively. A denotes the set of all the cropped images generated from D. Each cropped image $a_i \in \mathbf{A}$ is defined by $a_i = I_{f_i}(x_i, y_i, w_i, h_i)$ representing a pedestrian image of d_i, where I_{f_i} denotes the f_i frame of whole sequence. L denotes the index set representing IDs of pedestrians. Each detection vector d_i is labeled by one $l_i \in \mathbf{L}$. Then we can define $t_i \in \mathbf{T}$ as $t_i = (x_i, y_i, w_i, h_i, s_i, f_i, l_i)$, where $\mathbf{T}$ is the target set. In addition, we can augment t_i as $t_i = (x_i, y_i, w_i, h_i, s_i, f_i, l_i, z_i)$, where z_i denotes the deep re-identification feature of t_i defined by $z_i = NET_{reid}(a_i)$.

In this chapter, we propose a new multi-pedestrian tracking method based on trajectory matching based on appearance in dense scene settings. With discriminative

appearance features extracted by deep re-identification network, the robust trajectories can be generated by the appearance matching. To generate robust trajectories, the proposed method follows divide and conquer strategy. For details, whole image sequence $\mathbf{I}$ are divided into small sequences as $\mathbf{I} = \{I_1, I_2, ..., I_{N_{seq}}\}$, where N_{seq} denotes the maximum number of the divided sequences. Each $I_k \subset \mathbf{I}$, where I_k is a set containing the consequence $|I_k|$ images, we obtain detection results by using pedestrian detector and extract deep features by using deep neural network learned by person re-identification. Note that learning of person re-identification network similar to the learning of coarse-to-fine classification in that it classifies between pedestrian classes. So well-learned re-identification network is a good feature extractor for distinguishing different pedestrians. Then we can obtain $D_k \subset \mathbf{D}$ by pedestrian detector, which means the detection vector set obtained from I_k. Adding $l_i = 0$ and z_i obtained from re-identification network, we can obtain a tracking target set of I_k, T_k, which is defined by $T_k = \{t_1, t_2, ..., t_{n_k}\}$, where n_k is a maximum target number of T_k. At initialization of T_k, all the l_i of t_i in T_k have zero value, which means no ID assigned. Using T_k, we generate similarity matrix M to label the tracking targets, which is described in Sec. 5.3. Using M, we conduct two step target matching, appearance and distance based matching, to generate more robust trajectories. Here, we use the proposed matching method called by stable boundary selection (SBS) matching, which is described in Sec. 5.4. Then the short-term trajectories are generated as the results of SBS matching. Detection restoration for missed detection and the smoothing of the trajectories are applied to the short-term trajectories (described in Sec. 5.5 and Sec. 5.6). In next second stage of divide and conquer strategy, we generate the trajectory similarity matrix using the short-term trajectories. In a fixed N_m merging windows, which means $J_l = \{I_k, I_{k+1}, ..., I_{k+N_m-1}\}$, the goal of the second stage is to link the short-term trajectories by *hungarian* method to generate mid-term trajectories (described in Sec. 5.7). In final stage of divide and conquer strategy, we generate the trajectory similarity

matrix using the mid-term trajectories to make final long-term trajectories. Also this matching is solved by *hungarian* method on all the mid-term trajectories, which means final long-term trajectory set $K = \{J_l, J_{l+1}, ...\}$.

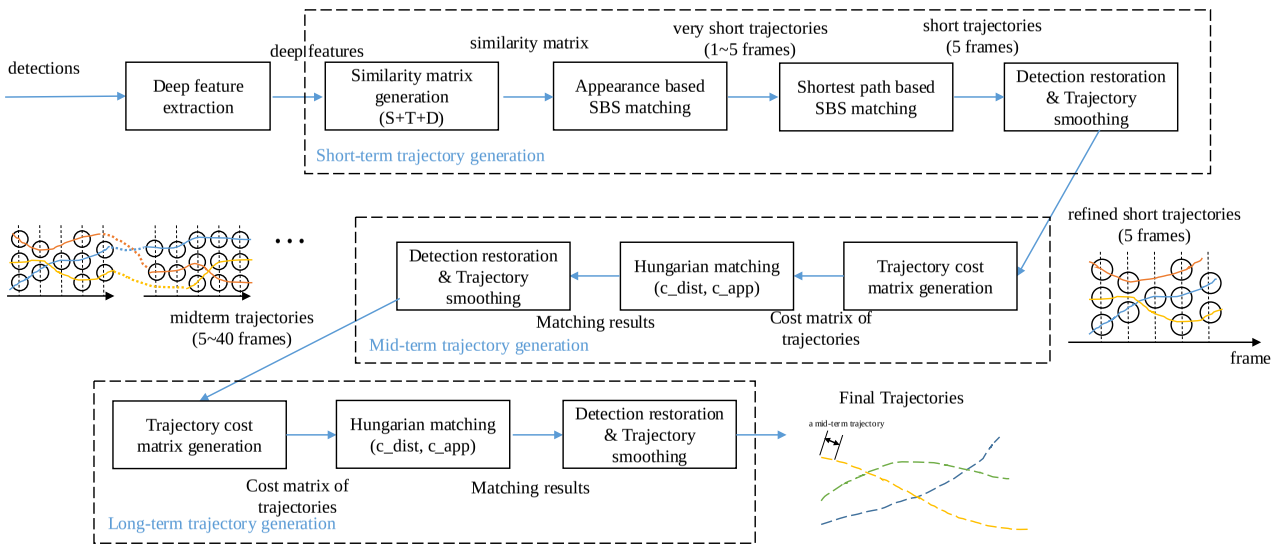


Figure 5.1: Overall scheme of the proposed method.

5.2 Similarity Matrix Generation

The goal of this step is to assign IDs to all targets. To assign IDs, we generate similarity symmetric matrix M_A by calculating the distance between all the t_i. The similarity symmetric matrix M_A is defined by

$$M_{A_{i,j}} = || \frac{z_i}{\sqrt{\sum z_i}} - \frac{z_j}{\sqrt{\sum z_j}} ||_2. \tag{5.1}$$

To prevent allocating the duplicated ID, we define tracking incompatibility matrix M_T defined by

$$\begin{gathered} M_{T_{i,j}} = c_{large} \quad \textit{if } f_i = f_j \\ 0 \quad \textit{else}, \end{gathered} \tag{5.2}$$

where c_{large} represent very large integer constant. Additionally it limits the pedestrian's box displacement to normal walking speed in I_k. This prevent the false matching of targets with large box distance. The box displacement matrix M_D is defined by

$$M_{D_{i,j}} = \sqrt{(x_i + \frac{w_i}{2} - x_j - \frac{w_j}{2})^2 + (y_i + h_i - y_j - h_j)^2}. \tag{5.3}$$

$M_{D_{i,j}}$ means the bottom centers displacement between t_i and t_j, which represent pedestrian 2-D location. But eq.5.2 is useless if frame interval between t_i and t_j is large. So eq.5.2 is complemented by

$$M_{D_{i,j}} = \frac{\sqrt{(x_i + \frac{w_i}{2} - x_j + \frac{w_j}{2})^2 + (y_i + h_i - y_j + h_j)^2}}{|f_i - f_j|}. \tag{5.4}$$

M_D is hinged to prevent the corruption of M_A as

$$\hat{M}_{D_{i,j}} = c_{large} \cdot max(0, M_{D_{i,j}} - c_{dis}), \tag{5.5}$$

where c_{dis} is a constant representing limitation of the box displacement. So we use $\hat{M}_D$ as M_D w.o.l.g. Finally we introduce the similarity matrix defined by

$$M = M_A + \lambda_T M_T + \lambda_D M_D. \tag{5.6}$$

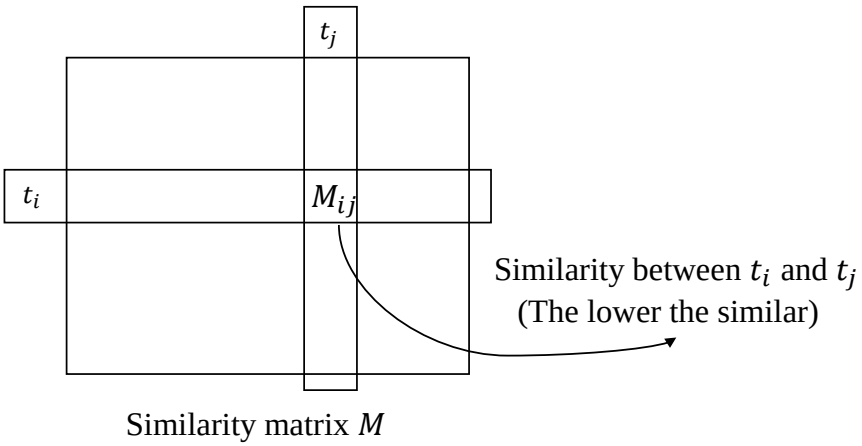


Figure 5.2: Similarity matrix between the tracking targets. This matrix is a symmetric and the value of the diagonal is zero.

5.3 Stable Boundary Selection

The goal of stable boundary selection (SBS) matching is to group and label the similar trajectories. Compared to *hungarian* matching, we design the SBS matching that can merge many trajectories to one trajectory, we call this trajectory merging. Since one short-term trajectory can be configured by many tracklets, the SBS matching is more reasonable than *hungarian* matching. In this section, we introduce a modified similarity matrix M_{t_i} for a specific target t_i. Before define M_{t_i}, we introduce a ordered vector set defined by

$$\begin{aligned} v_{t_i} &= \{M_{i,1}, M_{i,2}, ..., M_{i,n_k}\} \\ v'_{t_i} &= \{M_{i,b_{t_i}(1)}, M_{i,b_{t_i}(2)}, ..., M_{i,b_{t_i}(n_k)}\}, \end{aligned} \tag{5.7}$$

where v_{t_i} is the i-th row vector of M and v'_{t_i} is a modified vector of v_{t_i} ordered by the index set b_{t_i} (n_k is the number of row or column elements), which denotes the ascend sort of t_i. According to b_{t_i}, we define a modified similarity matrix M_{t_i} by defining the

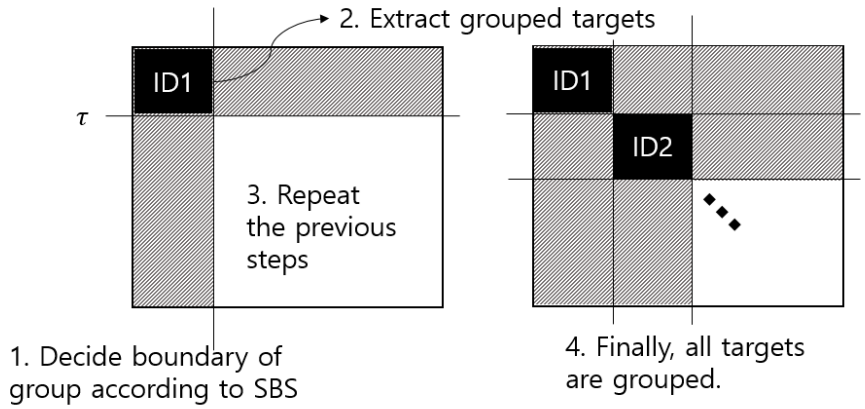


Figure 5.3: Iterative performing of SBS matching in the similarity matrix.

element of M_{t_i} as

$$M_{t_i}(j,k) = M(b_{t_i}(j), b_{t_i}(k)). \tag{5.8}$$

The basic principle of SBS matching is that similar targets of t_i tend to stick together in ordered similarity matrix M_{t_i}. Our goal is to group the similar targets of t_i in M_{t_i} and re-define v'_{t_j}, b_{t_j} and M_{t_j} to group the next similar targets. When grouping the similar targets of t_i, we need criteria to group. So we define a boundary difference for a criteria to group, which is defined by

$$\begin{aligned} c_l &= \frac{1}{l-1}\Big(\sum_{j=1}^{l-1}\sum_{k=1}^{l-1} M_{t_i}(j,k) - c_{l-1}\Big) \\ c_l &= \frac{1}{l-1}\sum_{j=1}^{l-1} M_{t_i}(i, b_{t_i}(j)). \end{aligned} \tag{5.9}$$

If boundary difference is large, the current group contains a dissimilar target. And if boundary difference is small, the current group maintain a stable boundary and contains the targets similar to t_i. We select the stable boundary by the criteria defined

by

$$c_l = \frac{1}{l-1} \sum_{j=1}^{l-1} M_{t_i}(i, b_{t_i}(j)) < \tau, \tag{5.10}$$

where τ is a constant denotes limit of boundary difference. This step is iterative performed until all the targets are grouped. This is described in Fig. 5.3. After all the targets are grouped, we assign the label to l_i of the targets meaning the ID of multi-pedestrian tracking.

5.4 Trajectory Smoothing

There are several reason that trajectory smoothing process are needed. First, pedestrian walks naturally under normal circumstances so the trajectory of the pedestrian is smooth. Second, the trajectory may contains the outliers due to the false measurement of the position (the coordinates of foot points or the size of bounding boxes) of some pedestrian. These factors cause significant tracking errors, but a trajectory smoothing algorithm can alleviate these problems. In this dissertation, we introduce two trajectory smoothing, smoothing based on the bounding box size and smoothing based on the location difference. The smoothing algorithm based on the size of the bounding box is based on the assumption that the size of the bounding box in the trajectory does not change significantly in a short frame. If the size of the target bounding box changes significantly in the trajectory, the smoothing algorithm modifies the bounding box properties (position, width and height). The criteria for modifying the bounding box properties is defined by

$$\begin{aligned} \bar{w_i} &= |w_i - w_{med}| \\ \bar{h_i} &= |h_i - h_{med}| \\ \tau_{size} &= max(0, \bar{w_i} - c_w) + max(0, \bar{h_i} - c_h), \end{aligned} \tag{5.11}$$

where w_{med} and h_{med} are the median value of width and height in the trajectory and c_w and c_h are the constant denoting allowable value changes. If τ_{size} is larger than zero, the smoothing algorithm modifies the bounding boxes as,

$$\begin{aligned} \tilde{w_i} &= w_{med} \\ \tilde{h_i} &= h_{med} \\ \tilde{x_i} &= x_i + \frac{(w_i - w_{med})}{2} \\ \tilde{y_i} &= y_i + (h_i - h_{med}). \end{aligned} \tag{5.12}$$

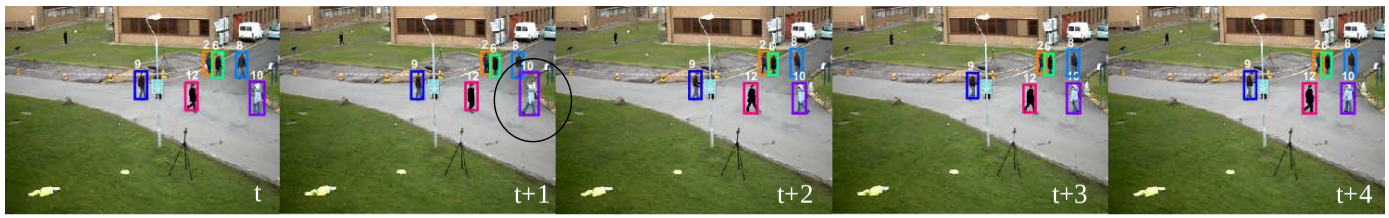
Instead of using the median of width and height as new values, the mean or other statistic can also used for the new width and height. However, the median value shows the best tracking performance, so we choose it. Although all the size of the bounding boxes in the trajectory is stable, the bottom center of some bounding box may be the outliers of the other bottom centers in the trajectory. We correct the outliers by using the smoothing algorithm. Here we assume that the $T_k^{l_j} = \{t_1, t_2, ...\}$ is the set having particular label l_j. Then the criteria τ_{loc} for modifying the bottom center of the bounding box is defined by

$$\begin{aligned} \bar{d_{loc}} &= \sqrt{((x_i + \frac{w_i}{2}) - (x_{i-1} + \frac{w_{i-1}}{2}))^2 + ((y_i + h_i) - (y_{i-1} + h_{i-1}))^2} \\ \tau_{loc} &= max(0, \bar{d_{loc}} - c_{loc}), \end{aligned} \tag{5.13}$$

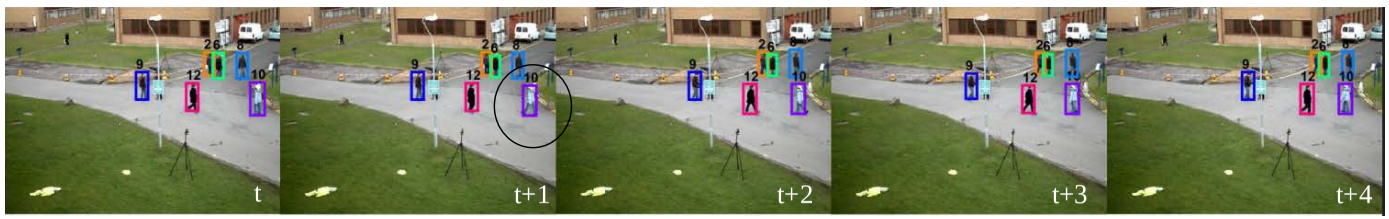
where c_{loc} is the constant denoting allowable value changes. If τ_{loc} is larger than zero, the smoothing algorithm modifies the bounding boxes as,

$$\begin{aligned} \tilde{w_i} &= \frac{r_{length} w_r + l_{length} w_l}{r_{length} + l_{length}} \\ \tilde{h_i} &= \frac{r_{length} h_r + l_{length} h_l}{r_{length} + l_{length}} \\ \tilde{x_i} &= \frac{r_{length}(x_r + w_r/2) + l_{length}(x_l + w_l/2)}{r_{length} + l_{length}} - \tilde{w_i}/2 \\ \tilde{y_i} &= \frac{r_{length}(y_r + h_r) + l_{length}(y_l + h_l}{r_{length} + l_{length}} - \tilde{h_i}, \end{aligned} \tag{5.14}$$

where (x_l, y_l, w_l, h_l) and (x_r, y_r, w_r, h_r) are the bounding boxes of the nearest unchanged target, which is the nearest boxes of previous and next frame respectively (r_{length} and l_{length} is the frame difference between current target and the nearest unchanged target). Modified bounding box's width, height and xy-coordinates are the inner points of the left and right boxes.



(a) Before adapting the smoothing algorithm based on the size change of bounding box



(b) After adapting the smoothing algorithm based on the size change of bounding box

Figure 5.4: Smoothing algorithm.

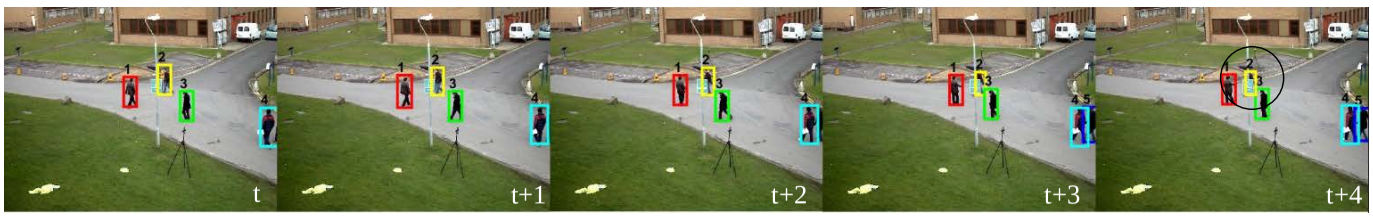
5.5 Detection Restoration

In this section, we introduce a method for the detection restoration to find the missed targets due to occlusion issues. In real application, pedestrian detector does not provide the perfect detection results. However, human can perceive the missed pedestrian visible through the next frame from the previous frame. Detection restoration is a replacement for this ability that humans can. Our matching based trajectory generation can match the trajectories even if the frame difference is large. There are two condition that detection restoration performing, 1) The frame difference between the two trajectories is not too large and 2) the movement speed does not exceed the average human speed. We divide the whole frame set in the trajectory into two disjoint frame sets, occupied frame set and empty frame set. The occupied frame set is a index set of the frame where the target box is on the trajectory. The empty frame set is an index set of frames where the target box is not in the trajectory. If the trajectory satisfying above two trajectory conditions, the detection restoration algorithm generate the new target bounding boxes in the empty frame set. The new target bounding boxes that the algorithm restore defined by

$$\begin{aligned}
\tilde{w}_{new} &= \frac{r_{length} w_r + l_{length} w_l}{r_{length} + l_{length}} \\
\tilde{h}_{new} &= \frac{r_{length} h_r + l_{length} h_l}{r_{length} + l_{length}} \\
\tilde{x}_{new} &= \frac{r_{length}(x_r + w_r/2) + l_{length}(x_l + w_l/2)}{r_{length} + l_{length}} - \tilde{w}_{new}/2 \\
\tilde{y}_{new} &= \frac{r_{length}(y_r + h_r) + l_{length}(y_l + h_l}{r_{length} + l_{length}} - \tilde{h}_{new},
\end{aligned} \tag{5.15}$$

where (x_l, y_l, w_l, h_l) and (x_r, y_r, w_r, h_r) are the bounding boxes of the nearest target in occupied frame set, which is the nearest boxes of previous and next frame respectively (r_{length} and l_{length} is the frame difference between current target and the nearest target in occupied frame set).

(a) Before adapting the detection restoration algorithm



(b) After adapting the detection restoration algorithm

Figure 5.5: detection restoration.

5.6 Trajectory Merging Process

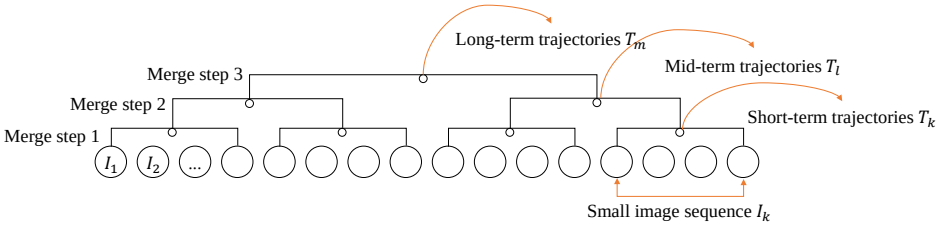


Figure 5.6: Trajectory merging process.

In this section, we introduce the trajectory merging process using divide and conquer strategy. As depicted in Fig. 5.6, the trajectory merging process conduct through three stage to generate short-term, mid-term and long-term trajectories respectively. The most important thing in generating the short-term trajectories is that it should be robust to ID switch. Since short-term trajectories are the most base trajectories of the merging process, the trajectories containing ID switch leads to false matching in merge step 2 and 3. Since this false matching is irrecoverable, we design the similarity matrix with hard constraints as shown in Eq. 5.6 in Sec. 5.3 and propose the SBS matching method as shown in Sec. 5.4. Once obtaining the robust short-term trajectories, mid-term and final trajectories can be more simply matched by using *hungarian* method. When matching between trajectories, the trajectory similarity is measured by the average of the target similarity. So we define a trajectory similarity matrix M_{T_i,T_j} between trajectory T_i and T_j defined by

$$M_{T_i,T_j}(k,l) = ||\frac{z_k^{T_i}}{\sqrt{\sum z_k^{T_i}}} - \frac{z_l^{T_j}}{\sqrt{\sum z_l^{T_j}}}||_2, \tag{5.16}$$

where $z_k^{T_i}$ and $z_l^{T_j}$ are the feature vectors of k-th target in T_i and l-th target in T_j, respectively. Now we define a cost matrix for the *hungarian* method to match the

trajectories in adjacent two image set I_i and I_j, which is defined by

$$M_{I_i,I_j}(k,l) = \frac{1}{N_{T_k} N_{T_l}} \sum_{m}^{N_{T_k}} \sum_{n}^{N_{T_l}} M_{T_k,T_l}(m,n), \tag{5.17}$$

where N_{T_k} and N_{T_l} is the number of targets in T_k and T_l, respectively.

Chapter 6

Experiments

6.1 Dataset and Evaluation Metric

6.1.1 Trajectory Matching in Overlapping Camera Settings

Dataset. The proposed network was evaluated on PETS 2009 dataset [45], and SNUPIL dataset [31]. In PETS 2009 dataset, we chose S2.L1 scenario for evaluation. The camera calibration information was given using *Tsai* camera calibration model [46].

Network Training. First we fine-tuned D-net to caltech pedestrian benchmark dataset [47]. We used pre-trained model [43] for feature extraction network. Then we followed the training setting and network parameters to train D-net like in [42]. After that, we trained L-net with detection proposals and corresponding 2-D PGPs in PETS 2009 and SNUPIL. In PETS 2009, we used 6, 8 views for training and 1, 5, 7 views for test. In SNUPIL, we used 4 views for training and 1, 2, 3 views for test. We used log loss for softmax classification for D-net and L-1 loss for regression task L-net. The detail setting for using multi-task loss is followed by [42].

Ground Truth. We generated about 30,000 samples for training L-net in PETS 2009

and 50,000 in SNUPIL. We generated 3-D ground truth trajectory using the pedestrian locations of each camera views. 3-D ground truth trajectory was estimated by minimizing the reconstruction error of 2-D ground truth based on Equation (4.6). Ground truth of 2-D PGP is generated every 5 frames in both dataset.

Measure. To evaluate the accuracy of PGP, we used the average of the Euclidean distance to ground truth as a measure. Specifically, we compute the Euclidean distance between the output of L-net and the ground truth. The detection box is determined based on 0.5 threshold intersection of union (IOU) based on the ground truth box. To evaluate tracking performance, we used the widespread CLEAR measures used in [48], called Multiple Object Tracking Accuracy (MOTA) and Multiple Object Tracking Precision (MOTP). And like in [49], we also used the metrics of Identity Switches (IDS), Track Fragments (FM), Mostly Tracked (MT), Partially Tracked (PT). We evaluate tracking performance every 5 frames in both dataset.

6.1.2 Trajectory Matching in Non-overlapping Camera Settings

Dataset. CUHK01 [50] contains 3,884 images of 971 identities. The images of each identity are captured from two disjoint cameras on the CUHK campus. The dataset provides only human-annotated bounding boxes. We split 971 identities into 871 training identities and 100 test identities. CUHK03 [12] contains 13,164 images of 1,360 identities. The images of each identity are captured from two disjoint cameras on the CUHK campus. The dataset provides both manually labeled bounding boxes and the bounding boxes obtained by a Deformable Part-based Model (DPM)-detector [51]. We experiment with our re-ranking algorithm on "labeled" data. We use 1,160 of the 1,360 identities for training, 100 identities for validations, and 100 identities for testing. MARKET1501 [52] consists of 32,668 bounding boxes of 1,501 identities generated by a DPM-detector on videos from six cameras. We use 751 identities for training and 750 identities for testing. DUKE [18, 53] dataset is generated by the images cap-

tured by eight cameras. Training and test sets both consist of 702 identities. The dataset provides manually annotated bounding boxes.

Measure. We adopt a widely used evaluation protocol, where Re-ID performance is measured by cumulative matching characteristics (CMC) [12] and mean average precision (mAP) for Re-ID [52].

Network parameters. We embed 128-dimension features by adding two additionally fully-connected layers to the ResNet-50 network [54]. In ResNet-50 network, We deleted the last fully-connected layer (fc-layer), and the two fc-layers are added in the order of size 2048×1024 and 1024×128. We set the epoch to 150, momentum to 0.9, batch size to 32 and initial learning rate to 3×10^{-4} with a polynomial decaying. We use Adam-optimizer [55] with parameters ϵ=0.001, β_1=0.9, β_2=0.999. In training the network, we used triplet loss on CUHK01 and CUHK03, whereas we used identification loss [18] on MARKET1501 and DUKE.

6.1.3 Robust Trajectory Matching in Dense Scene Settings

Dataset. We experiment the proposed multi-pedestrian tracking algorithm on the public tracking dataset, PETS2009 and 2D-MOT15. PETS, the IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, has provided the datasets containing tracking, detection and people counting. Among the PETS datasets, PETS2009 dataset is one of the most popular tracking dataset nowaday. There are three sequences for tracking scenario with different people density, which are S2L1 (low density), S2L2 (high density) and S2L3 (very high density). Among the three sequences, we choose the S2L1 scenario as our base dataset for tuning parameters and ablative studies. MOTChallenge, The Multiple Object Tracking Benchmark, provide several detection and tracking dataset containing 2D-MOT15, 3D-MOT15, MOT16, MOT17Det, MOT17, MOT19Det, MOT19, MOT20Det and MOT20. In tracking datasets of MOTChallenge, they provide the results of the public pedestrian detector, Aggregated Channel

Feature (ACF) [2], Deformable Part Model (DPM) [1]and Regionlets [56]. Therefore, the proposed method can be fairly compared with other multi-pedestrian tracking algorithms with fixed pedestrian detector. MOTChallenge also provide development kits for evaluating the tracking performance, we use the evaluation code of that. There are many tracking scenario in MOTChallenge. Among them, we choose 2D-MOT15 because it is most widely used dataset and contains various sub-dataset compared to previous datasets. Among them, we choose PETS2009-S2L1, TUD-Stadmitte, TUD-Campus, Venice-2 and ADL-Rundle-6, which are the sub-dataset of 2D-MOT15. For fair competition, they provide the detection results of ACF [2] so we use that for our algorithm and the others as the input detection boxes.

Measure. To evaluate tracking performance, we used the widespread CLEAR measures used in [48], called Multiple Object Tracking Accuracy (MOTA) and Multiple Object Tracking Precision (MOTP) also explained in Sec 6.1.1. In this section, we describe more details of the metrics. FN_t denotes the number of false negatives at t-th frame, which means the tracker missed the targets. FP_t denotes the number of false positives at t-th frame, which means the tracker find the non-human objects. $IDSW_t$ denotes the number of identity switches at t-th frame and GT_t is the number of ground truth in t-th frame. Then MOTA is defined by

$$MOTA = \frac{\sum_t (FN_t + FP_t + IDSW_t)}{\sum_t GT_t}. \tag{6.1}$$

Also MOTP is defined by

$$MOTP = \frac{\sum_i matcher(t_i)}{\sum_f m_f}, \tag{6.2}$$

where $matcher$ is the function gives one if the input target is matched with ground truth with overlap ratio greater than 0.5 (if not gives zero) and m_f denotes the number of matches in frame f. We also consider the track quality measures, which are mostly tracked (MT), partially tracked (PT) and mostly lost (ML). If a trajectory is included

in MT, it is successfully tracked for at least 80% of its ground truth trajectory. Here, ID-switch does not affect on the performance of MT. If a trajectory is included in ML, it is partially tracked for less than 20% of its ground truth trajectory. If a trajectory is not included both in MT and ML, we determine the trajectory is included in PT. In addition, we compute the identification precision (IDP), identification recall (IDR) and identification F_1 score IDF_1. MOTA and MOTP are metrics that are not related to identity retention, so we need metrics to measure how well tracking IDs are preserved. To compute these measure, we compute $IDFN$ (number of missing ID), $IDFP$ (number of false assignmnt of ID)and $IDTP$ (number of well assigned ID).

$$\begin{aligned} IDP &= \frac{IDTP}{IDTP + IDFP} \\ IDR &= \frac{TDTP}{IDTP + IDFN} \\ IDF_1 &= \frac{2IDTP}{2IDTP + IDFP + IDFN}. \end{aligned} \tag{6.3}$$

Parameters. We determine all parameters by the experimental experience. We use the feature network as the person re-identification network, which is trained on Market-1501 dataset with Resnet-50. Note that there are no fine-tuning or parameter tuning of the feature network. We modulate the number divided sequences N_{seq} by setting $|I_k|= 5$. Also we set $N_{w2} = 8$ the number of sequences in one merging window, which means that the trajectories generated on 40 sequential frames are in one merging window. $c_{large} = 10^{10}$, $c_{dis} = 25$, $\lambda_T = c_{large}$ and $\lambda_D = 1$, respectively. In SBS matching, the constant denoting the limit of the boundary difference is set as $\tau = 0.45$. In trajectory smoothing, we set $c_w = 10$, $c_h = 15$ and $c_{loc} = 7$.

6.2 Results and Discussion

6.2.1 Trajectory Matching in Overlapping Camera Settings

Localization Accuracy. The quantitative results regarding the accuracy of pedestrian localization are depicted in Table 6.2. In most cases, DL-net shows the best performance. Additionally, we evaluate D-net, removing L-net from DL-net. D-net is similar to a fast rcnn network [42]. As shown in Table 6.1, L-net has the positive effect of improving 2-D PGP accuracy by adding it to D-net.

MCMTT Performance. The quantitative evaluation of the tracking performance is depicted in Table 6.1. The overall tracking performance of the proposed method is better than those associated with other detection results. In PETS 2009, the proposed method shows a much better performance than other algorithms. In MOTP, in particular, representing the accuracy on 3-D localization is greatly improved. This implies that the accurately estimated 2-D PGP has a positive effect on the generation of the accurate position in the 3-D trajectory. As shown in Table 6.2, the tracking result with DPM shows a high MOTA, while D-net shows a high MOTP. DPM shows a high recall performance in the detection of the pedestrian, and D-net shows a low localization error. In addition, DL-net shows a higher MOTA performance than D-net. This shows that an accurate 2-D PGP can decrease the ambiguity between pedestrians who are close together. Since DPM missed distorted pedestrians, DPM in SNUPIL does not yield a high MOTA as in PETS 2009. However DL-net's 2-D PGPs are robust to image distortion problem. As a result, DL-net shows high MOTA and MOTP in SNUPIL.

6.2.2 Trajectory Matching in Non-overlapping Camera Settings

6.2.2.1 Ablation Study

We carried out ablation experiments on three candidate sets and the first responsible gallery set (FRGS). The results are depicted in Fig.3. All the six variants use the same

Dataset	Detector	Cameras	MOTA [%]	MOTP [%]	MT	PT	FM	IDS
PETS 2009 S2.L1	HOG-SVM [5]	1,5,7	77.16	41.75	19	4	28	12
	DPM [1]	1,5,7	97.34	59.57	23	0	2	4
	D-Net	1,5,7	90.36	80.13	23	0	2	4
	DL-Net	1,5,7	**97.59**	**80.37**	23	0	**1**	**3**
SNUPIL	DPM [1]	1,2,3	53.48	69.90	4	6	11	18
	D-Net	1,2,3	74.38	78.63	8	2	10	16
	DL-Net	1,2,3	**81.09**	**82.92**	8	2	**4**	**5**

Table 6.1: Tracking performance evaluation with different detection methods in PETS 2009 and SNUPIL

Method	CAM 1	CAM 5	CAM7
HOG-SVM [5]	20.20	25.28	27.50
DPM [1]	9.92	13.87	14.43
D-net	8.17	11.16	13.69
DL-Net	**7.07**	**10.01**	**11.59**

Table 6.2: Quantitative results on the SNUPIL dataset. The error shown in the table is an euclidean distance between grounding position and ground truth. We used a pixel unit and 0.25 intersection of union (IOU) constant to determine the boxes corresponding to ground truth.

deep-feature extracted from Resnet-50. “rrs-p”, “rrs-g”, and “rrs-pg” in Fig. 3 represent three configurations that use only probe set, only gallery set, and probe + gallery set, respectively. “w.o. frgs” is a variant of “rrs-pg” where FRGS is removed. “krecip” represents a re-ranking method using normal KNN. We measured the mean and standard deviation of mAP by repeating the above methods 100 times on CUHK01.

From the ablation experiments, we obtained the following empirical observations. First, as shown in the result of “rrs-g” and “rrs-pg”, the gallery set greatly contributes

Table 6.3: Performance comparison on CUHK01 and CUHK03 among the re-ranking methods. The table shows the average of rank accuracies and mAPs. Bold indicates the best one.

Method	CUHK01		CUHK03	
	R-1	mAP	R-1	mAP
DeepReID (CVPR14)	27.9	-	-	-
IDLA (CVPR15)	65.0	-	-	-
PersonNet (ArXiv16)	71.1	-	-	-
SIR-CIR (CVPR16)	71.8	-	-	-
Deep Metric (ECCV16)	69.4	-	-	-
DCSL (IJCAI16)	89.6	-	80.2	-
Deep Part (ICCV17)	88.5	-	85.4	**90.9**
Resnet50	86.7	86.6	79.8	77.8
Resnet50+ECN (CVPR18)	81.5	83.4	83.5	84.2
Resnet50+Krecip (CVPR17)	88.1	89.8	85.7	87.4
Resnet50+RRS	**90.7**	**91.6**	**88.3**	89.9

to the performance improvement. Second, FRGS, induced from the the gallery rank list, also makes a significant contribution to performance. This implies that the gallery rank list is important as much as the probe rank list. Third, the ordered KNN based method (“rrs-pg”) outperforms the normal KNN based method (“krecip”). This means that the ordering of the shared neighbors in KNN takes an important role to distinguish the hard cases.

6.2.2.2 Performance Evaluation

In this subsection, we present the test results on CUHK01, CUHK03, Market1501 and DUKE dataset. In CUHK01 and CUHK03 experiment, we randomly choose 100

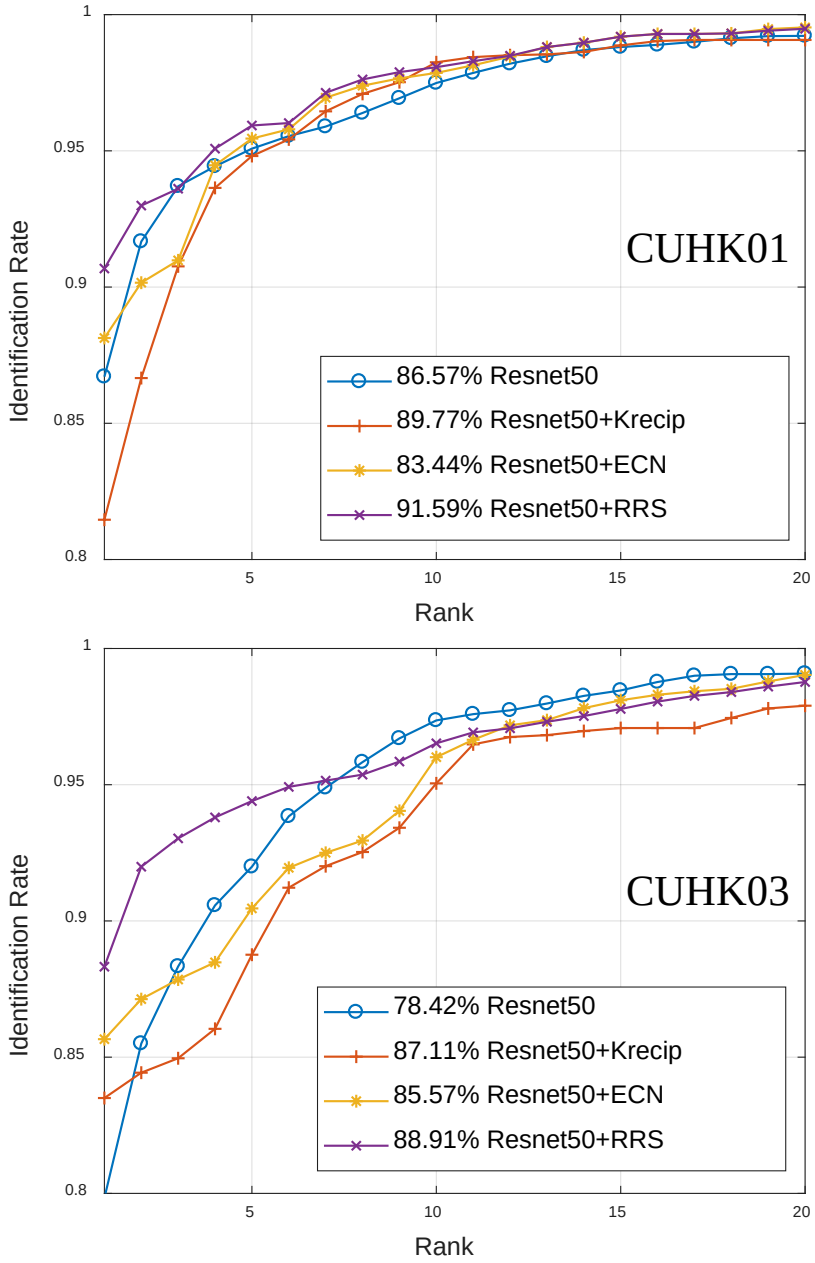


Figure 6.1: CMC-curves on CUHK01 and CUHK03. mAP values are given in the box.

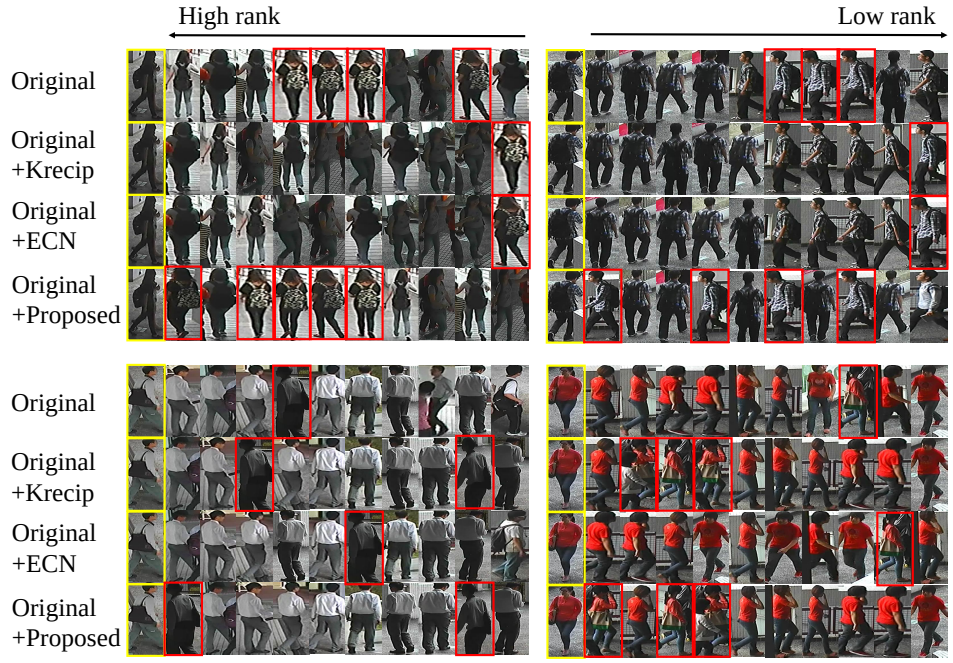


Figure 6.2: Qualitative comparison of appearance ambiguity cases on CUHK03. The three re-ranking methods (Krecip, ECN, proposed) are compared with the baseline network denoted 'Original' (without re-ranking). A yellow box indicates a probe image and a red box indicates the true gallery with the same ID as the probe. As depicted in the pictures, compared with the existing re-ranking methods, the proposed re-ranking method tends to arrange the true gallery to the front-ranked positions.

test identities and repeat the experiment 100 times to obtain the mean values of CMC and mAP. For evaluation on CUHK01 and CUHK03, the following methods were used: "DeepReID" [12], "IDLA [57]", "PersonNet" [58], "SIR-CIR" [59], "Deep Metric" [60], "DCSL" [61], "Deep Part" [38], "Krecip" [26], "ECN" [27]. Table 1 shows the average values of rank$(1, 5, 10, 20)$-accuracy and mAP for the CUHK01 test set. On CUHK01, The the proposed re-ranking method shows the state-of-the-art performance, 90.7% rank-1 accuracy and 91.6% mAP. On CUHK03, the proposed re-ranking method shows the best rank-1 accuracy (88.3%, 2.6 higher than the sec-

Table 6.4: Performance comparison on Market1501 and DUKE. The table shows the average of rank accuracies and mAPs. Bold indicates the best one.

Method	Market-1501		DUKE	
	R-1	mAP	R-1	mAP
ACRN (CVPR17-W)	83.6	62.6	72.6	52.0
SVD (ICCV17)	82.3	62.1	76.7	56.8
Part Aligned (ICCV17)	81.0	63.4	-	-
PDC (ICCV17)	84.1	63.4	-	-
JLML (IJCAI17)	85.1	65.5	-	-
DPFL (ICCV17-W)	88.6	72.6	79.2	60.6
PSE+ECN (CVPR18)	90.4	84.0	85.2	79.8
HA-CNN (CVPR18)	91.2	75.7	80.5	63.8
Resnet50	89.4	73.2	80.1	63.0
Resnet50+Krecip (CVPR17)	91.5	**87.1**	84.7	79.8
Resnet50+RRS	**92.0**	**87.1**	**85.6**	**79.9**

ond best), and mAP(89.9%). Among the re-ranking methods, the proposed re-ranking method shows best Re-ID performance. Also, in Fig. 4, we depict Cumulative Matching Characteristics(CMC)-curve of re-ranking methods from rank-1 to rank-20. As shown in Fig. 4, "Resnet50+RRS" shows the best identification rate until rank-7 and also shows the best mAP.

In addition, we conduct experiments on all the test identities in Market1501 and DUKE to obtain the rank accuracy and mAP. In evaluating Re-ID performance on Market1501and DUKE, all the images in gallery set are used in all literature and the standard deviation is zero. Hence we omitted the standard deviation in Table 2. In baseline, ID discriminative embedding [18] was used for training the network (denoted "IDE" in the table). For evaluation on Market1501 and DUKE, the following methods

Method	IDF1 [%]	IDP [%]	IDR [%]	MT	PT	ML	FP	FN	IDs	FM	MOTA [%]	MOTP [%]
$M_A + \lambda_T M_T + \lambda_D M_D$	83.8	83.0	84.6	18	1	0	381	297	10	56	84.6	71.9
$M_A + \lambda_D M_D$	86.3	85.5	87.2	18	1	0	383	295	5	57	84.7	71.9
$M_A + \lambda_T M_T$	77.7	76.8	78.6	18	1	0	406	304	10	56	83.9	72.0
M_A	78.6	77.4	79.7	18	1	0	440	310	13	58	83.0	71.9
w.o. smooth wh	83.2	82.5	84.0	18	1	0	406	322	10	61	83.5	71.9
w.o. smooth loc	83.4	82.6	84.1	18	1	0	402	318	11	89	83.7	71.6
w.o. dr	81.9	84.1	79.8	15	4	0	333	562	16	137	79.6	72.1
w.o. smooth wh+loc	83.2	82.4	84.0	18	1	0	411	327	10	92	83.3	71.6
w.o. smooth wh+dr	81.4	83.6	79.3	15	4	0	355	584	16	137	78.7	72.1
w.o. smooth loc+dr	81.2	83.4	79.1	15	4	0	355	584	23	162	78.5	71.7
w.o. smooth wh+loc+dr	81.2	83.4	79.1	15	4	0	354	583	24	156	78.5	71.7

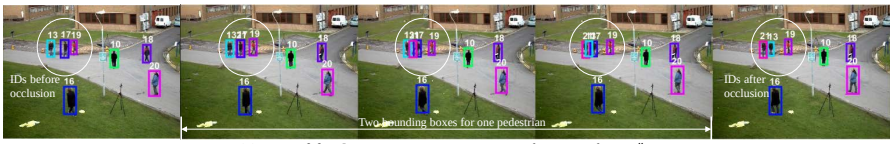
Table 6.5: Ablative study on PETS2009 S2L1. The red number denotes the best and the blue one denotes second for each tracking metrics.

were used: "ACRN" [62], "SVD [63]", "Deep Part" [38], "PDC" [64], "JLML" [65], "DCSL" [61], "DPFL" [66], "PSE+ECN" [27], "HA-CNN" [21]. These results are in Table 2. The proposed re-ranking method shows the best rank-1 accuracy (92.0%, 85.6%) and mAP (87.1%, 79.9%) on Market1501 and DUKE dataset. The qualitative comparison is given in Fig. 6.2.

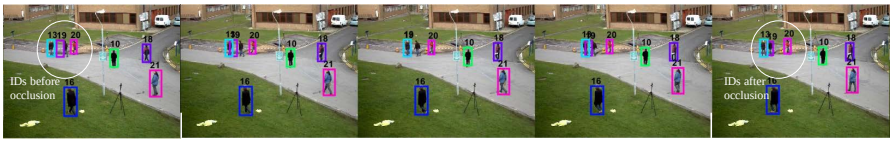
6.2.3 Robust Trajectory Matching in Dense Scene Settings

6.2.3.1 Ablative Studies

First, we introduce our ablative studies for the various settings. In Table 6.5, "$M_A + \lambda_T M_T + \lambda_D M_D$" is the base model of the proposed tracking algorithm. "$M_A + \lambda_T M_T$", "$M_A + \lambda_D M_D$" and "M_A" are the modified version of the model by extracting M_D, M_T and both M_D and M_T, respectively. "w.o.smooth" denotes the modified model by removing smoothing algorithm from the base model. "w.o.smooth wh" denotes the model that is removed smoothing based on the bounding box size. "w.o.smooth loc" denotes the model that is removed smoothing based on the location



(a) Part of final trajectories using "$M_A + \lambda_T M_T + \lambda_D M_D$"



(b) Part of final trajectories using "$M_A + \lambda_D M_D$"

Figure 6.3: The examples of ID switches occurred by M_T. In (a), ID 17 in a white circle is changed to to ID 21 after 3 frames caused by duplicated detection boxes on ID 17. In (b) without M_T, there are not ID switches but some misalignment of ID 19 box.

(a) Final trajectories without detection restoration

(b) Final trajectories with detection restoration

Figure 6.4: The effectiveness of the detection restoration Compared to the trajectories in (a), the trajectories in (b) retain the IDs well.

difference. “w.o.dr” denotes the model that is removed detection restoration.

First, we discuss about the effectiveness of each terms in similarity matrix. As shown in Table 6.5, we focus on the model “$M_A + \lambda_D M_D$” shows higher tracking performance than “$M_A + \lambda_T M_T + \lambda_D M_D$”,which is especially robust to ID switches. The role of M_T is to prevent the false matching between the targets having same time

stamps. However, some times this constraints causes ID switches when there are two bounding boxes for one pedestrian. Then the trajectory may be broken at that point and new trajectory with new id is generated. This situation is described in Fig. 6.3. In general case, it is not allowed that one ID is duplicated appeared in one frame. We choose a lower performance but more reasonable model "$M_A + \lambda_T M_T + \lambda_D M_D$". We can also observe that "$M_A + \lambda_T M_T$" and "M_A" shows tracking performance drops. This means that M_D prevents the false matching between the targets showing large box displacements.

Second, we discuss about the effectiveness of the smoothing algorithm and detection restoration. Compared to the base model, the false positives are increased with "w.o smooth wh". This implies that the smoothing algorithm increases the overlap ratio between ground truth and the modified detection boxes, and this leads to decrease the false positives. "w.o smooth wh+loc" shows also performance drop so we argue that our two version of smoothing algorithms both contribute to increase the tracking performance. We also see that detection restoration also affect on the false positives and false negatives by observing the results of "w.o smooth wh+loc+dr" and "w.o smooth dr". By restoring the missed detection boxes, the number of false negatives are effectively dropped.

6.2.3.2 Quantitative Results

In this section, we compare the proposed method to other tracking methods. We compare the proposed method to four multi-object tracking algorithms. The first algorithm is the method called by Continuous Energy Minimization (CEM) [4], which proposed various energy functions representing occlusion, trajectory maintenance and ID switches. The second algorithm is the method called by Similar Multi-Object Tracking (SMOT) [67], which uses motion dynamics to distinguish targets with similar appearance. The third algorithm is the method called by TBD [68], which present a proba-

Method	IDF1 [%]	IDP [%]	IDR [%]	MT	PT	ML	FP	FN	IDs	FM	MOTA [%]	MOTP [%]
Proposed	**83.8**	**83.0**	**84.6**	**18**	1	0	381	**297**	**10**	56	**84.6**	71.9
CEM	66.7	66.7	66.7	**18**	1	0	**368**	367	30	**39**	82.9	**73.1**
SMOT	-	-	-	12	7	0	322	1154	99	204	66.1	71.6
TBD	-	-	-	12	7	0	707	870	239	126	60.9	71.2
LP2D	-	-	-	18	1	0	2984	278	207	89	22.5	70.9

Table 6.6: Comparison results on PETS2009 S2L1

Method	IDF1 [%]	IDP [%]	IDR [%]	MT	PT	ML	FP	FN	IDs	FM	MOTA [%]	MOTP [%]
Proposed	**64.7**	76.3	**56.1**	4	6	0	**44**	349	**7**	15	**65.4**	66.1
CEM	64.5	**82.0**	53.1	**5**	4	1	45	452	**7**	**6**	56.4	65.4
SMOT	-	-	-	4	6	0	62	334	16	26	64.4	**70.2**
TBD	-	-	-	8	2	0	196	**205**	28	13	62.9	69.5
LP2D	-	-	-	6	4	0	1144	216	42	10	-21.2	64.8

Table 6.7: Comparison results on TUD-Stadmitte

Method	IDF1 [%]	IDP [%]	IDR [%]	MT	PT	ML	FP	FN	IDs	FM	MOTA [%]	MOTP [%]
Proposed	**58.1**	67.5	**51.0**	4	3	1	17	105	**3**	**4**	**65.2**	73.2
CEM	55.8	**73.0**	45.1	1	6	1	13	150	7	7	52.6	72.3
TBD	-	-	-	**5**	3	0	43	**85**	9	12	61.8	**74.9**
SMOT	-	-	-	1	5	2	**3**	189	11	12	43.5	74.7
LP2D	-	-	-	1	7	0	132	173	30	20	6.7	69.4

Table 6.8: Comparison results on TUD-Campus

bilistic generative model for multi-object tracking to estimate 3D scene layout, location and orientation of objects in the scene. The last algorithm is the MOT baseline called

Method	IDF1 [%]	IDP [%]	IDR [%]	MT	PT	ML	FP	FN	IDs	FM	MOTA [%]	MOTP [%]
Proposed	**36.5**	**43.7**	**31.4**	**5**	17	4	**1750**	3766	57	137	**22.0**	**74.0**
CEM	33.5	43.6	29.9	4	16	6	1890	4144	**42**	**52**	14.9	72.6
LP2D	-	-	-	**5**	18	3	7400	**2988**	267	132	-49.2	71.9

Table 6.9: Comparison results on Venice-2

Method	IDF1 [%]	IDP [%]	IDR [%]	MT	PT	ML	FP	FN	IDs	FM	MOTA [%]	MOTP [%]
Proposed	**42.1**	**56.3**	**33.6**	0	**20**	4	**669**	2693	**51**	95	**31.9**	71.4
CEM	36.5	50.0	28.5	1	18	5	744	2898	53	**45**	26.2	**72.5**
LP2D	-	-	-	**4**	17	3	4256	**1938**	329	144	-30.2	69.4

Table 6.10: Comparison results on ADL-Rundle-6

by Linear programming on 2D image coordinates (LP2D) [69]. We experiments all the comparison on the same public pedestrian detector, ACF. The experimental results of the proposed methods on PETS2009 S2L1 dataset are shown in Table 6.6. The proposed method shows the best results on the metrics of IDF1, IDP, IDR, MT, FN, IDs and MOTA. The experimental results of the proposed methods TUD-Stadmitte dataset are shown in Table 6.7. The proposed method shows the best results on the metrics of IDF1, IDR, FP, IDs and MOTA. The experimental results of the proposed methods on TUD-Campus dataset are shown in Table 6.8. The proposed method shows the best results on the metrics of IDF1, IDR, IDs, FM and MOTA.

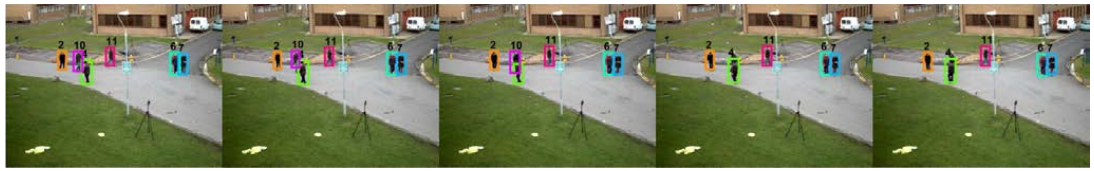
6.2.3.3 Qualitative Results

In this section, we propose the qualitative results of the proposed tracking algorithm. We provide three types of qualitative results, which are the results of the short-term

trajectories, mid-term trajectories and final long-term trajectories. Each row is listed in chronological order, meaning that the bounding boxes of the same color have the same ID. We can see robust short-term trajectories and use them for match to generate the trajectories of mid-term and long-term trajectories.

(a) Short-term trajectories for 301 to 305 frames

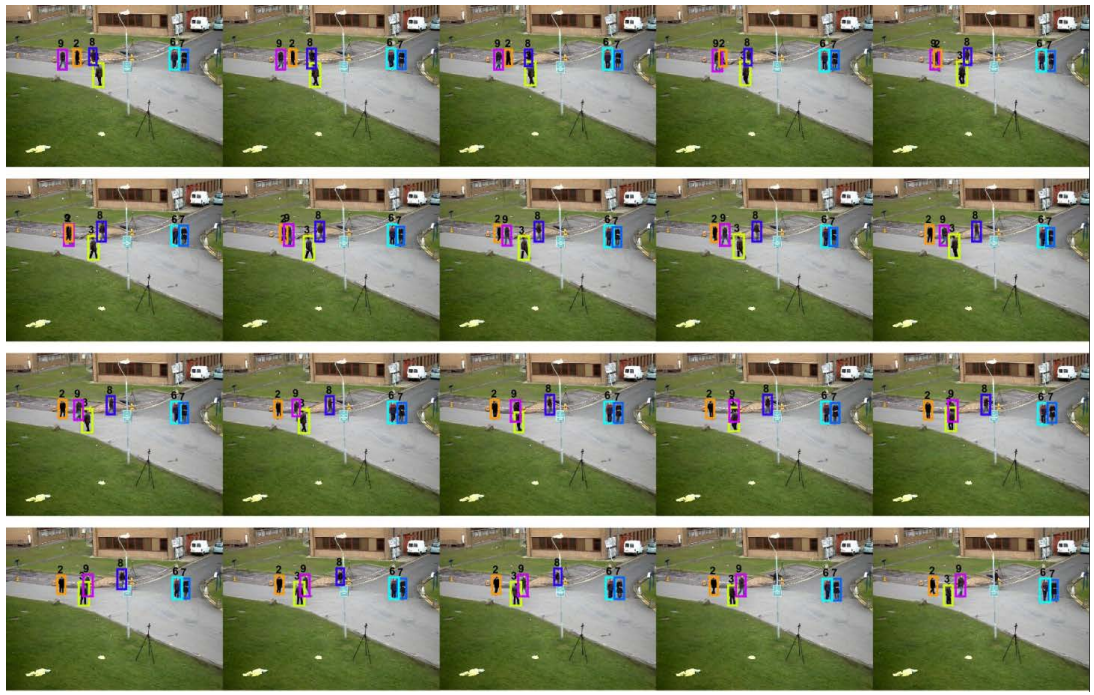
(b) Short-term trajectories for 306 to 310 frames



(c) Short-term trajectories for 311 to 315 frames

(d) Short-term trajectories for 316 to 320 frames

Figure 6.5: The results of the short-term trajectories on PETS2009S2L1.



(a) Mid-term trajectories for 301 to 320 frames

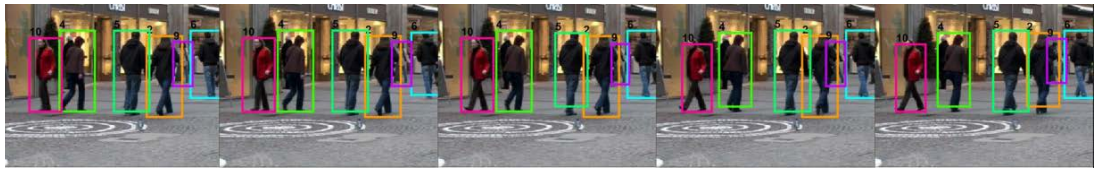
Figure 6.6: The results of the mid-term trajectories on PETS2009S2L1.

(a) long-term trajectories for 301 to 320 frames

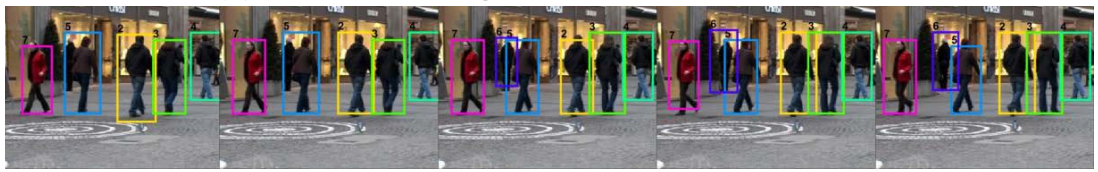
Figure 6.7: The first results of the long-term trajectories on PETS2009S2L1.

(a) long-term trajectories for 321 to 340 frames

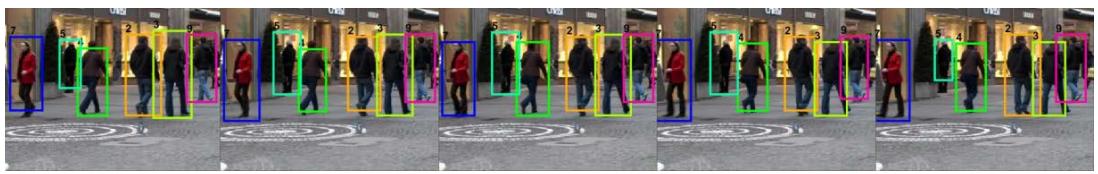
Figure 6.8: The second results of the long-term trajectories on PETS2009S2L1.



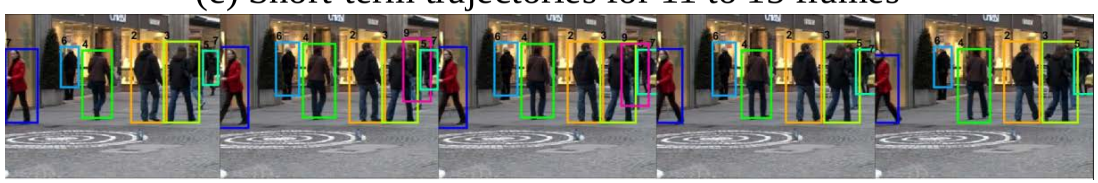
(a) Short-term trajectories for 1 to 5 frames



(b) Short-term trajectories for 6 to 10 frames

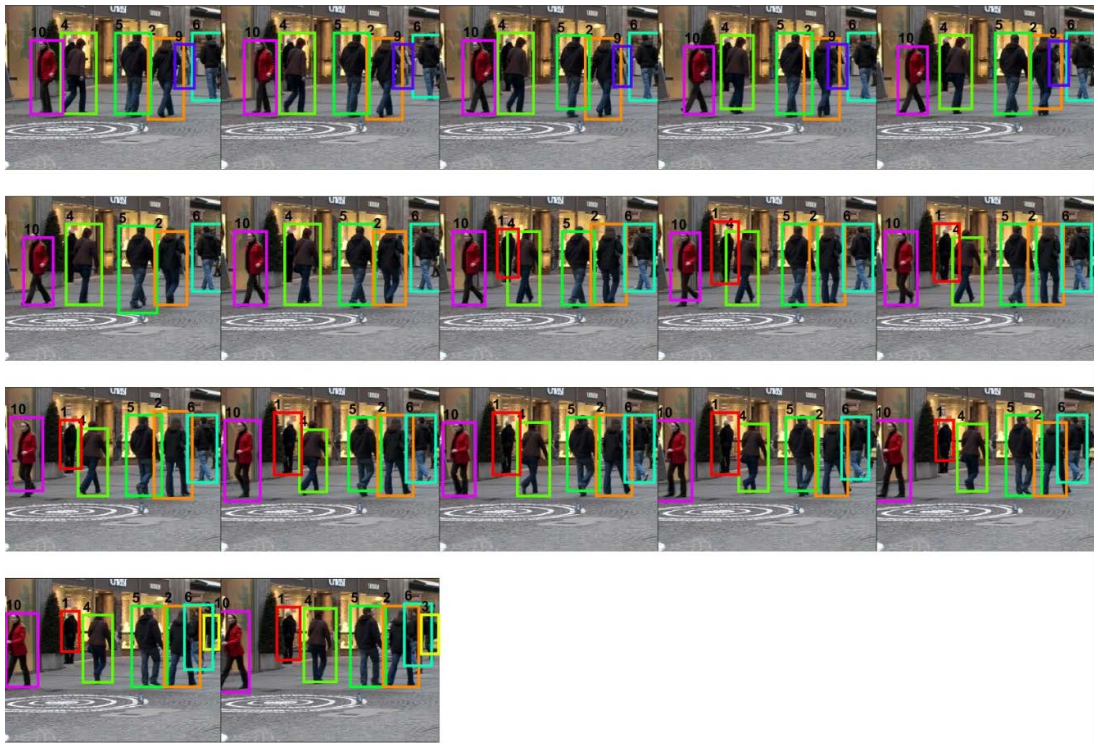


(c) Short-term trajectories for 11 to 15 frames



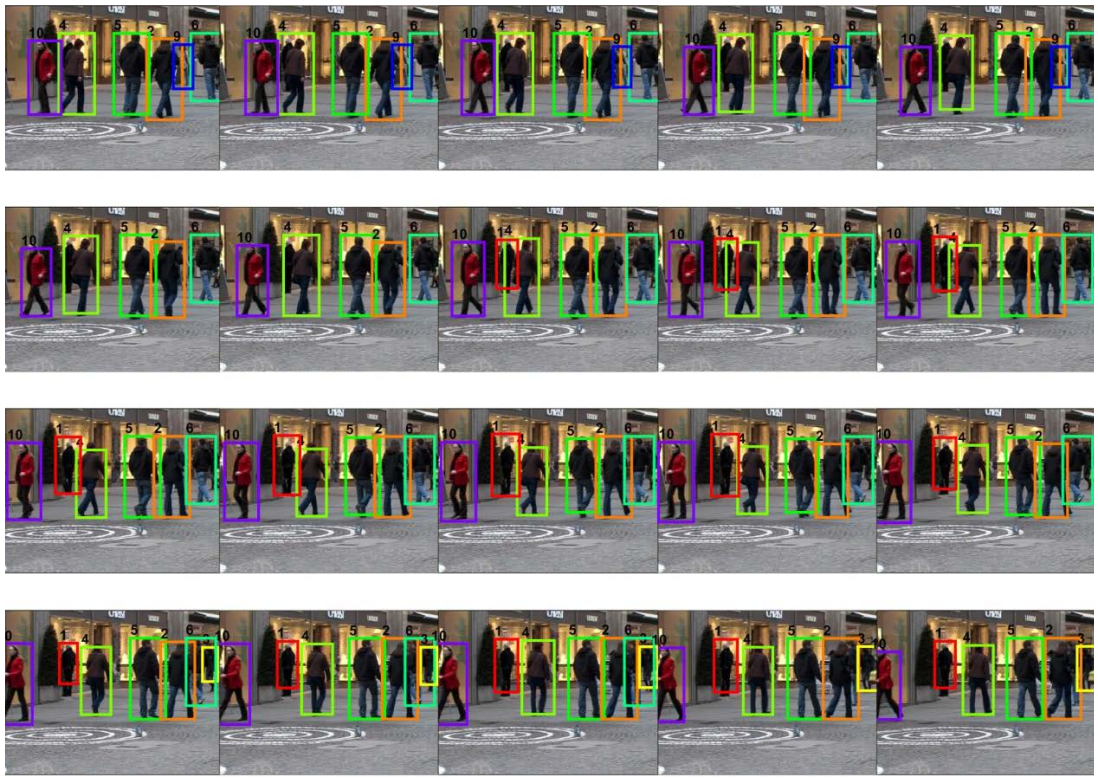
(d) Short-term trajectories for 16 to 20 frames

Figure 6.9: The results of the short-term trajectories on TUD-Stadmitte.



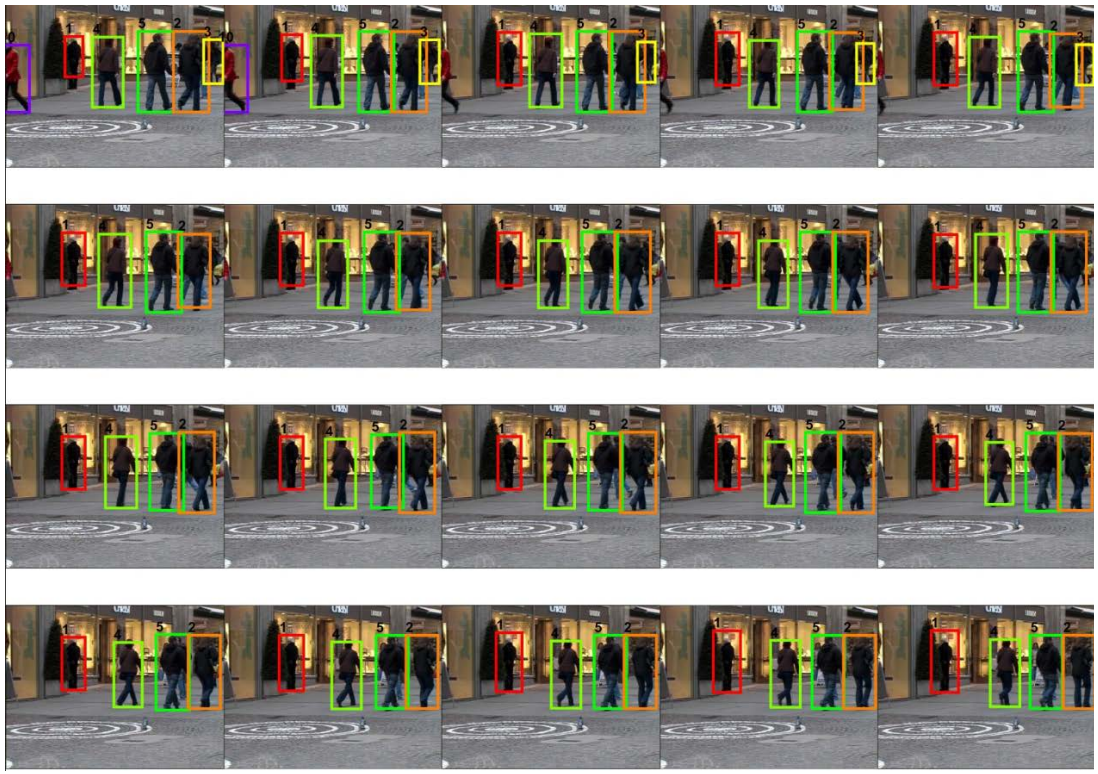
(a) Mid-term trajectories for 1 to 17 frames

Figure 6.10: The results of the mid-term trajectories on TUD-Stadmitte.



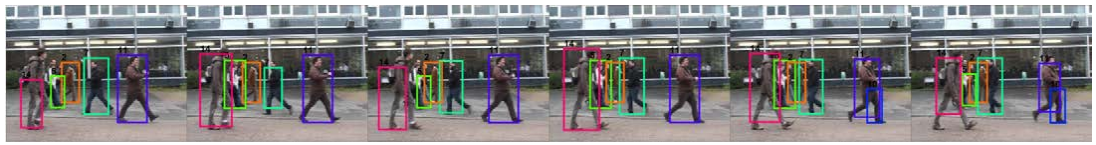
(a) Long-term trajectories for 1 to 20 frames

Figure 6.11: The first results of the long-term trajectories on TUD-Stadmitte.

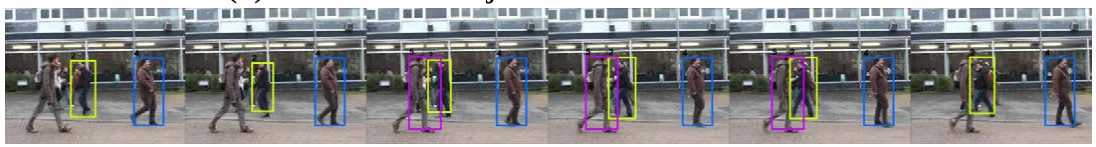


(a) Long-term trajectories for 21 to 40 frames

Figure 6.12: The second results of the long-term trajectories on TUD-Stadmitte.



(a) Short-term trajectories for 1 to 6 frames

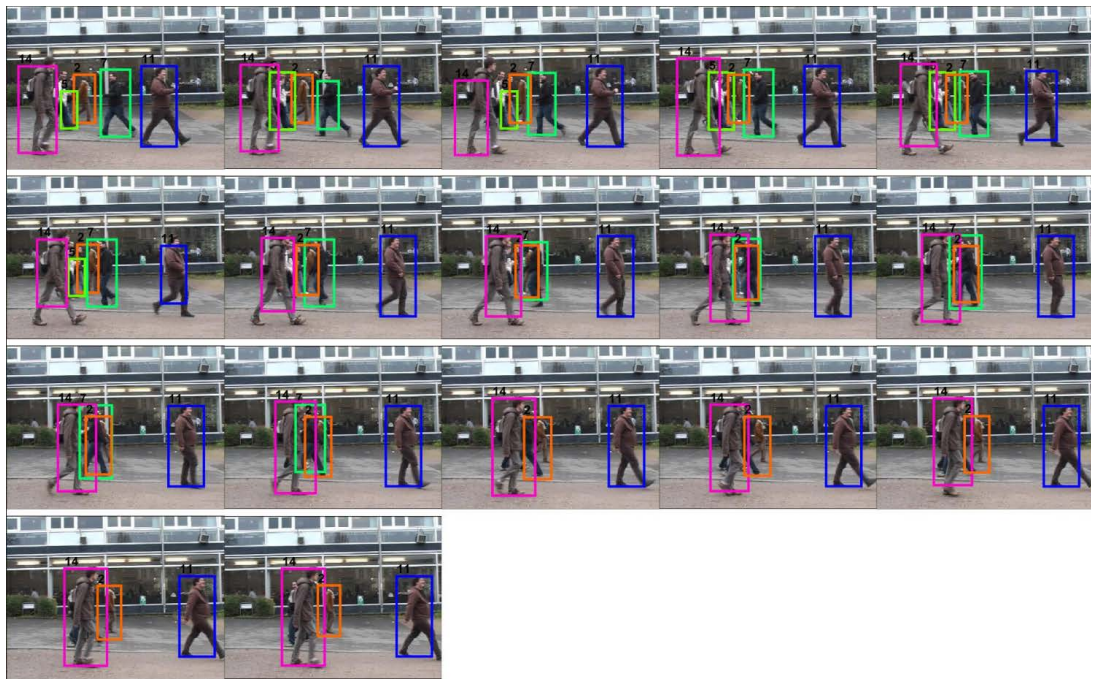


(b) Short-term trajectories for 7 to 12 frames

(c) Short-term trajectories for 13 to 18 frames

(d) Short-term trajectories for 19 to 24 frames

Figure 6.13: The results of the short-term trajectories on TUD-Campus.

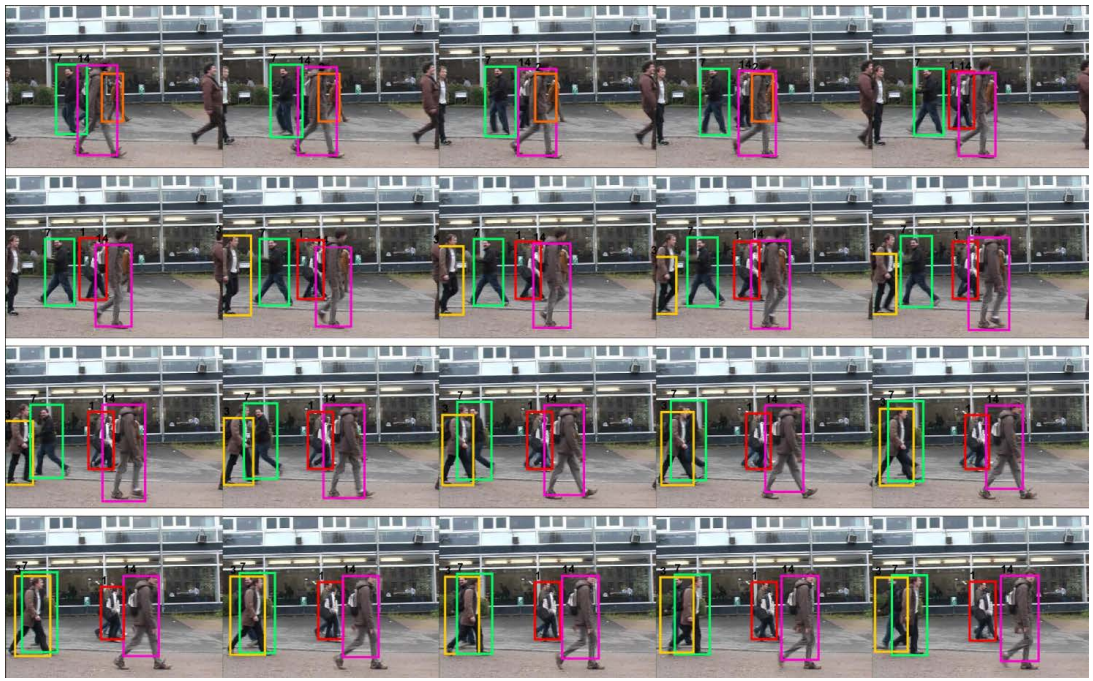


(a) Mid-term trajectories for 1 to 17 frames

Figure 6.14: The results of the mid-term trajectories on TUD-Campus.

(a) long-term trajectories for 1 to 20 frames

Figure 6.15: The first results of the long-term trajectories on TUD-Campus.



(a) long-term trajectories for 21 to 40 frames

Figure 6.16: The second results of the long-term trajectories on TUD-Campus.

Chapter 7

Conclusions and Future Works

7.1 Concluding Remarks

In this Dissertation, we proposed the algorithms for wide-area multi-pedestrian tracking. First, we have proposed a deep-learning network that simultaneously detects pedestrians and estimates their ground positions. Compared to previous tracking methods, ours has focused on the importance of pedestrian localization accuracy. Unlike the method of box regression, the proposed method directly gives a 2-D ground position from the detection box to improve localization accuracy. It is meaningful that enhancing the localization accuracy of detections significantly improves the performance of MOTP, representing both tracking and localization in 3-D space. Second, a novel re-ranking method has been proposed to improve the accuracy in person re-identification. In the proposed re-ranking method, three key factors contribute to the accuracy improvement. The first factor is the ranking-reflected similarity (RSS) between the ordered set of K-nearest neighbors (OKNN) of a probe and that of a gallery. The second factor is the selection of candidate neighbor sets to construct the OKNNs. The third

factor is the re-ranking procedure, where priority is given to the galleries likely to have the same ID as the given probe rather than re-ranking the entire galleries. As validated in the experiments, the three factors in the proposed re-ranking method lead to the improvement of Re-ID accuracy, outperforming the state-of-the-art re-ranking methods. Finally, we proposed a wide-area multi-pedestrian tracking framework based on hierarchical trajectory matching. In the proposed tracking method, we generate the trajectories by the strategy of divide and conquer method. From the generation of the short-term trajectories, we propose a novel deep-feature matching method called stable boundary selection (SBS). In SBS matching, the detections are clustered by the group similarity of deep features, so that robust short-term trajectories can be generated. The mid-term and long-term trajectories are generated by matching the short-term and mid-term trajectories using *hungarian* method. With two smoothing algorithms and detection restoration algorithm, the proposed tracking method shows the state-of-the-art tracking accuracy in three public tracking dataset.

7.2 Future Works

Remained part of our research is to define the unified framework for integrating pedestrian detection, person re-identification(Re-ID) and multi-pedestrian tracking. In recent years, some researchers have studied about joint learning of pedestrian detection, person Re-ID, and multi-pedestrian tracking.

Regarding the related works, in 2017, [70] proposed a joint detection and identification feature learning. They integrated pedestrian detection and Re-ID in one network. The goal of this network is to find and identify a person at one stage. They called this work as *person search*. Otherwise, in 2018, [71] proposed the framework to extract features for multi-target multi-camera tracking and re-identification. They utilize the Re-ID network to generate a trajectory of person. After that, they used correlation clustering to associate the generated trajectories to one cluster.

The main objective of our future work is to unify a detection, Re-ID and MCMTT in one framework. Fig. 7.1 represents our overall framework. In our framework, image $t-1$ and t are fed to the network input. After passing the region proposal network, the candidate boxes are generated of image $t-1$ and t. Using ROI-pooling, the images from the candidate boxes are fed to detection network. And we can obtain the scores of the boxes and the 2-D position of grounding points(PGP). 2-D PGP is used for calculating the reconstruction cost of tracking loss discussed in Chapter 2.2 Otherwise, the candidate boxes are also fed to Re-ID network using selective ROI-pooling. selective ROI-pooling is used to provide only human images to Re-ID network. Comparing two features, the Re-ID network presents the similarity score between two boxes. This score is used to calculate pairwise cost of tracking loss function.

In T frames, the network is iterative propagation the above step.

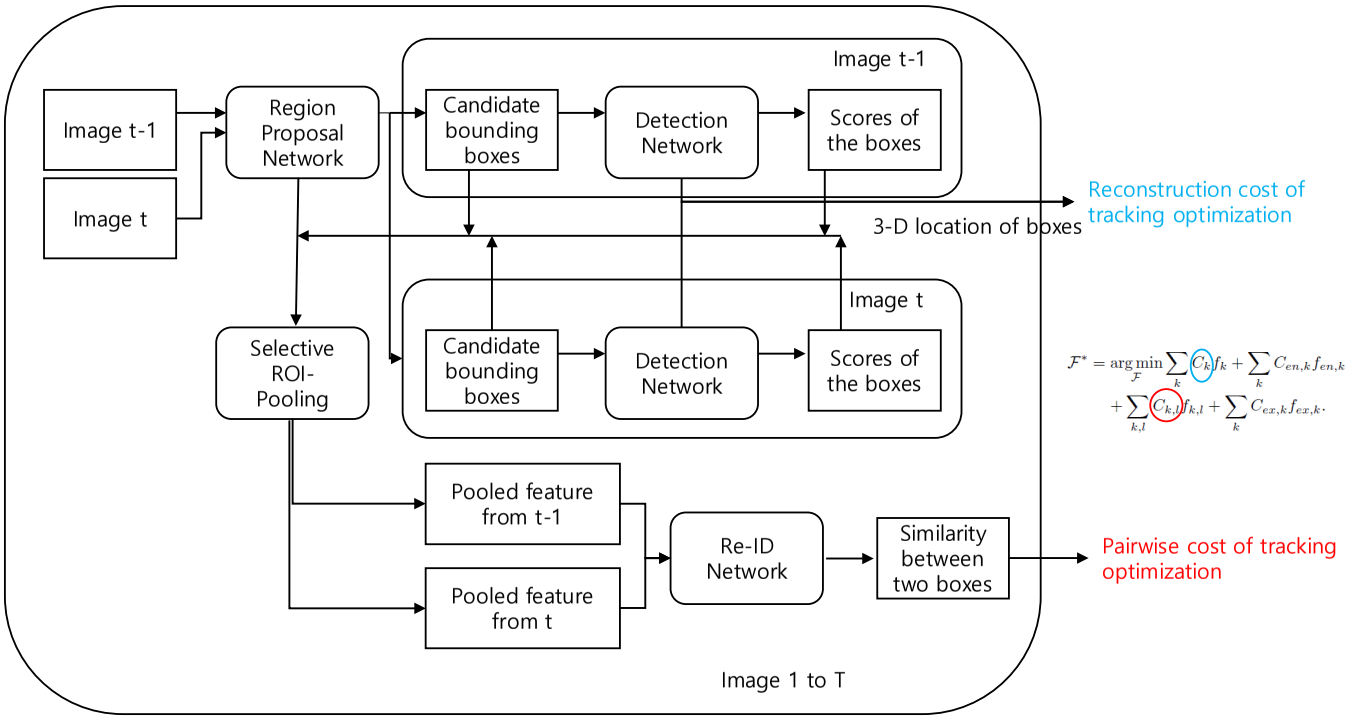


Figure 7.1: The joint framework of pedestrian detection, person re-identification and multi-pedestrian tracking.

After the tracking loss is calculated, the tracking error is back-propagated to the network. The learning schedule is as follows:

1. Using all the images and ground truth of bounding boxes, the detection network is initially trained, meanwhile the weights of Re-ID network are fixed.

2. After initializing the detection network, Re-ID network and detection network are jointly trained.

3. We divide a video to sub-videos having T frames. Images of each sub-videos are fed to the network for back-propagating tracking error.

Since we have studied the pedestrian detection and Re-ID, we will focus mainly on the association of tracking part with the remained part. The main target of our framework will be dataset which consists of non-overlapped static cameras. The goal of this work is to ensure that the unified framework outperforms other single frameworks.

Bibliography

[1] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Computer Vision and Pattern Recognition (CVPR), 2008 IEEE Conference on*. IEEE, 2008, pp. 1–8.

[2] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 8, pp. 1532–1545, 2014.

[3] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua, "Multiple object tracking using k-shortest paths optimization," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 9, pp. 1806–1819, 2011.

[4] A. Milan, S. Roth, and K. Schindler, "Continuous energy minimization for multitarget tracking," *IEEE TPAMI*, vol. 36, no. 1, pp. 58–72, 2014.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition (CVPR), 2005 IEEE Conference on*, vol. 1. IEEE, 2005, pp. 886–893.

[6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[7] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Computer Vision and Pattern Recognition (CVPR)*, 2013.

[8] ——, "Person re-identification by salience matching," in *International Conference on Computer Vision (ICCV)*, 2013.

[9] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, "Local fisher discriminant analysis for pedestrian re-identification," in *Computer Vision and Pattern Recognition (CVPR)*, 2013.

[10] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification," in *Computer Vision and Pattern Recognition (CVPR)*, 2014.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[12] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Computer Vision and Pattern Recognition (CVPR)*, 2014.

[13] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *International Conference on Pattern Recognition (ICPR)*, 2014.

[14] R. O. Duda, P. E. Hart, and D. G. Stork, "Pattern classification and scene analysis," 1973.

[15] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," in *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[16] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: a deep quadruplet network for person re-identification," in *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[17] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[18] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," *arXiv preprint arXiv:1701.07717*, 2017.

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[20] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[21] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[22] V.-H. Nguyen, T. D. Ngo, K. M. Nguyen, D. A. Duong, K. Nguyen, and D.-D. Le, "Re-ranking for person re-identification," in *Soft Computing and Pattern Recognition (SoCPaR)*, 2013.

[23] L. An, X. Chen, M. Kafai, S. Yang, and B. Bhanu, "Improving person re-identification by soft biometrics based reranking," in *International Conference on Distributed Smart Cameras (ICDSC)*, 2013.

[24] J. Garcia, N. Martinel, C. Micheloni, and A. Gardel, "Person re-identification ranking optimisation by discriminant context information analysis," in *International Conference on Computer Vision (ICCV)*, 2015.

[25] Q. Leng, R. Hu, C. Liang, Y. Wang, and J. Chen, "Person re-identification with content and context re-ranking," *Multimedia Tools and Applications*, vol. 74, 2015.

[26] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Reranking person reidentification with k-reciprocal encoding," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 2017.

[27] M. S. Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen, "A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking," *arXiv preprint arXiv:1711.10378*, 2017.

[28] Y. Rao, J. Lin, J. Lu, and J. Zhou, "Learning discriminative aggregation network for video-based face recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3781–3790.

[29] G. Chen, J. Lu, M. Yang, and J. Zhou, "Spatial-temporal attention-aware learning for video-based person re-identification," *IEEE Transactions on Image Processing*, 2019.

[30] M. Byeon, S. Oh, K. Kim, H.-J. Yoo, and J. Y. Choi, "Efficient spatio-temporal data association using multidimensional assignment in multi-camera multi-target tracking." in *BMVC*, 2015, pp. 68–1.

[31] H. Yoo, K. Kim, M. Byeon, Y. Jeon, and J. Y. Choi, "Online scheme for multiple camera multiple target tracking based on multiple hypothesis tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, 2016.

[32] M. Hofmann, D. Wolf, and G. Rigoll, "Hypergraphs for joint multi-view reconstruction and multi-object tracking," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, 2013, pp. 3650–3657.

[33] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multicamera people tracking with a probabilistic occupancy map," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 267–282, 2008.

[34] S. M. Khan and M. Shah, "Tracking multiple occluding people by localizing on multiple scene planes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 3, pp. 505–519, 2009.

[35] L. Leal-Taixé, G. Pons-Moll, and B. Rosenhahn, "Branch-and-price global optimization for multi-view multi-target tracking," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1987–1994.

[36] H. Possegger, S. Sternig, T. Mauthner, P. M. Roth, and H. Bischof, "Robust Real-Time Tracking of Multiple Objects by Volumetric Mass Densities," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, 2013.

[37] W. Nam, P. Dollár, and J. H. Han, "Local decorrelation for improved pedestrian detection," in *Advances in Neural Information Processing Systems*, 2014, pp. 424–432.

[38] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification." in *International Conference on Computer Vision (ICCV)*, 2017.

[39] J. Almazan, B. Gajic, N. Murray, and D. Larlus, "Re-id done right: towards good practices for person re-identification," *arXiv preprint arXiv:1801.05339*, 2018.

[40] R. A. Jarvis and E. A. Patrick, "Clustering using a similarity measure based on shared near neighbors," *IEEE Transactions on computers*, vol. 100, 1973.

[41] D. Qin, S. Gammeter, L. Bossard, T. Quack, and L. Van Gool, "Hello neighbor: Accurate object retrieval with k-reciprocal nearest neighbors," in *Computer Vision and Pattern Recognition (CVPR)*, 2011.

[42] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

[43] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[44] "Gurobi optimization," http://www.gurobi.com/downloads/gurobi-optimizer/.

[45] A. Ellis, A. Shahrokni, and J. M. Ferryman, "Pets2009 and winter-pets 2009 results: A combined evaluation," in *Performance Evaluation of Tracking and Surveillance (PETS-Winter), 2009 Twelfth IEEE International Workshop on*. IEEE, 2009, pp. 1–8.

[46] R. Tsai, "A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses," *IEEE Journal on Robotics and Automation*, vol. 3, no. 4, pp. 323–344, 1987.

[47] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *Pattern Analysis and Machine Intelligence*, vol. 34, 2012.

[48] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, and J. Zhang, "Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics,

and protocol," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 319–336, 2009.

[49] Y. Li, C. Huang, and R. Nevatia, "Learning to associate: Hybridboosted multi-target tracker for crowded scene," in *Computer Vision and Pattern Recognition (CVPR), 2009 IEEE Conference on*. IEEE, 2009, pp. 2953–2960.

[50] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *Asian Conference on Computer Vision (ACCV)*, 2012.

[51] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 32, 2010.

[52] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *International Conference on Computer Vision (ICCV)*, 2015.

[53] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking (ECCVW)*, 2016.

[54] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[55] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[56] X. Wang, M. Yang, S. Zhu, and Y. Lin, "Regionlets for generic object detection," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 17–24.

[57] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3908–3916.

[58] L. Wu, C. Shen, and A. v. d. Hengel, "Personnet: Person re-identification with deep convolutional neural networks," *arXiv preprint arXiv:1601.07255*, 2016.

[59] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang, "Joint learning of single-image and cross-image representations for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1288–1296.

[60] H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei, W. Zheng, and S. Z. Li, "Embedding deep metric for person re-identification: A study against large variations," in *European conference on computer vision*. Springer, 2016, pp. 732–748.

[61] Y. Zhang, X. Li, L. Zhao, and Z. Zhang, "Semantics-aware deep correspondence structure learning for robust person re-identification." in *IJCAI*, 2016, pp. 3545–3551.

[62] A. Schumann and R. Stiefelhagen, "Person re-identification by deep learning attribute-complementary information," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 20–28.

[63] Y. Sun, L. Zheng, W. Deng, and S. Wang, "Svdnet for pedestrian retrieval," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3800–3808.

[64] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3960–3969.

[65] W. Li, X. Zhu, and S. Gong, "Person re-identification by deep joint learning of multi-loss classification," *arXiv preprint arXiv:1705.04724*, 2017.

[66] Y. Chen, X. Zhu, and S. Gong, "Person re-identification by deep learning multi-scale representations," in *IEEE International Conference on Computer Vision*, 2017, pp. 2590–2600.

[67] C. Dicle, O. I. Camps, and M. Sznaier, "The way they move: Tracking multiple targets with similar appearance," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 2304–2311.

[68] A. Geiger, M. Lauer, C. Wojek, C. Stiller, and R. Urtasun, "3d traffic scene understanding from movable platforms," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 5, pp. 1012–1025, 2013.

[69] L. Leal-Taixé, M. Fenzi, A. Kuznetsova, B. Rosenhahn, and S. Savarese, "Learning an image-based motion context for multiple people tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3542–3549.

[70] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3415–3424.

[71] E. Ristani and C. Tomasi, "Features for multi-target multi-camera tracking and re-identification," 2018.

Abstract

광역 추적 문제의 목적은 시간 간격이나 사람 밀도에 관계없이 겹치거나 겹치지 않는 카메라에 나타나는 보행자를 추적하는 것이다. 단일 카메라 추적에서 감지 상자의 겹침을 사용하는 데이터 연결은 추적 문제를 해결하는 데 사용되지만 여전히 모양 모호성 문제가 있다. 그러나 광역 추적에는 감지 상자의 겹침을 사용하지 않고 사람의 외형 유사성에 중점을 둔 추적 체계가 필요하다. 이 논문에서는 광역 다중 보행자 추적 (WaMuPeT)에 대한 추적 체계를 제안한다. WaMuPeT를 달성하기 위해 겹치는 카메라 설정 (3 장), 겹치지 않는 카메라 설정 (4 장) 에서의 궤적 일치 그리고 빽빽한 장면 설정 (5 장)에서 강인한 궤적 일치에 대해서 제안한다.

겹치는 카메라 설정에서의 궤적 매칭 (3 장)에서는 여러 카메라를 사용하여 보행자를 정확하게 3D 지역화하고 추적하기위한 새로운 딥 러닝 아키텍처를 제안한다. 딥 러닝 네트워크는 감지 네트워크와 로컬라이제이션 네트워크의 두 가지 네트워크로 구성된다. 탐지 네트워크는 보행자 탐지를 제공하고 현지화 네트워크는 탐지 상자 내에서 보행자의 지상 위치를 추정한다. 또한 두 개의 네트워크를 효과적으로 연결하기 위해주의 패스 필터가 도입되었다. 두 네트워크에서 얻은 탐지 제안 및 2D 접지 위치를 사용하여 최소 비용의 네트워크 흐름 접근 방식을 통해 다중 카메라 다중 대상 3D 지역화 및 추적 알고리즘이 개발된다. 실험에서 제안 된 방법이 3D 지역화 및 추적 성능을 향상시키는 것으로 나타났다.

겹치지 않는 카메라 설정에서의 궤적 일치 (4 장)에서, 우리는 순위가 반영된 메트릭을 사용하여 두개의 순서가 지정된 K-최근 접 이웃 (OKNN) 세트 사이의 유사성을 측정한다. 순위 반영 유사성 (RSS)에 대해 제안 된 메트릭은 두 OKNN 사이의 공유 요소의 순위를 반영합니다. RSS를 사용하여, 순위 순서의 관점에서 프로브의 이웃과 유사한 이웃을 갖는 갤러리를 우선 순위 화하는 재순위 절차가 제안된다. 실험에서 제안 된 방법이 최신 방법에 추가되어 Re-ID 정확도가 향상됨을 보여준다.

고밀도 장면 설정에서 강력한 궤적 일치 (5 장)에서, 우리는 고밀도 장면에서 강력한 궤적을 생성하기 위해 다중 보행자 추적을 위한 새로운 프레임 워크를 제안한다. 제안된 추적 방법에서는 분할 및 정복 방법 전략에 따른 궤적 매칭을 기반으로 추적 방법을 제안한다. 이 전략에서, 단기, 중기 및 장기 궤적은 각각의 궤적 병합 단계에 의해 생성된다. 또한 SBS (Stable Boundary Selection)라는 새로운 기능 매칭 기법을 제안한다. SBS 매칭에서, 탐지는 깊은 특징의 그룹 유사성에 의해 군집화되어, 강력한 궤적이 생성 될 수 있다. 제안 된 추적 방법은 평활 알고리즘과 탐지 복원 알고리즘을 통해 3 개의 공개 추적 데이터 세트에서 최첨단 추적 정확도를 보여준다.

Keywords: 광역 추적, 다중 보행자 추적, 보행자 감지, 보행자 지역화, 개인 재 식별, 재 순위 지정

Student Number: 2013-20748